FACULTY OF LAW
Lund University

Teona Gelashvili

Hate Speech on Social Media:

Implications of private regulation and governance gaps

JAMM07 Master Thesis

International Human Rights Law
30 Higher Education Credits

Supervisor: Karol Nowak

Spring Term 2018

Table of contents

Summary

Hate speech on social media hardly remains unnoticed. Contents involving hateful messages vary from "kill a Jew day" to "kick a ginger day" and could target anyone irrespective of their status, identity, location and so forth. Even when hate speech is not materialized into a hate-motivated crime, the damage is done – victims are being labeled, marginalised and exposed to negative stereotyping. The overall consequences of online hate can be the dehumanisation of individuals or groups of individuals.

The need for proper strategies to tackle hate speech on social media is unquestionable. The core focus of the thesis is not to find a solution to the challenge, but rather to identify central problems that have contributed to the formation of the existing reality. To unrave the contributing factors, a holistic analysis of both international human rights principles regarding hate speech and the practical application of those standards is necessary.

Accordingly, this thesis examines whether the protections provided against hate speech are a sufficient response to the challenge arising from the specific nature of expression on social media. The distinctive characteristics of social media play a key role and provide an ideal venue to target and reach a wide audience across the globe.

Under those circumstances, the decisive role is not played by states or international institutions, but by social media platforms, that are predominantly private institutions. States, that are the central duty-bearers to respect, to protect and to fulfil human rights, in fact, have a very limited opportunity to influence the process of regulating expression on social media. This is due to a number of factors. Complications related to the technical infrastructure of social media platforms, applicable legislation, the definition of different concepts are a few examples of the factors that hinder advancement of protection of human rights. Furthermore, the unwillingness of states to collaborate is an additional factor that contributes to insufficient regulation of hate speech on social media and it is questionable whether a problem can be dealt with without collective efforts.

The conclusions drawn in this paper is the result of a critical assessment of the current international legal framework and the self-regulation mechanism adopted by the private actors. The disconnect between the international human rights framework and its implementation results in government gaps – a dangerous trend that can result in arbitrariness and selective application of different rules. Thus, it creates a risk of either excessive regulation, or leaving a significant part of society unprotected against exposure to hateful expression. This thesis further provides some proposals for the future consideration regarding the possible solution of the problem.

Keywords: hate speech, social media, freedom of expression, self-regulation, accountability of private actors, and privatization of censorship.

Acknowledgments

I am deeply grateful to the many people who have contributed to the development of this paper. First and foremost, I would like to express my deepest gratitude to my supervisor, Dr. Karol Nowak, for his comments, suggestions, criticism, and encouragement throughout the development of the thesis.

My special thanks are extended to the academic staff at Lund University and Raoul Wallenberg Institute of Human Rights, who made my experience truly valuable.

I have been very fortunate to be able to study on the Faculty of Law at Lund University, which would not be possible without the Visby scholarship from the Swedish Institute to pursue my master's degree at Lund.

I want further to thank to my classmates and students at Lund University for inspiration, sharing knowledge and experience. Special thanks to my classmate and friend - Dureti Fulas for being so supportive, encouraging and kind. I would like to further express gratitude to my opponent, Raul Gerardo Rodriguez Quintana for his valuable comments on the defense seminar.

Finally, I would like to thank my family and friends, especially my parents - Marine and Simon for their continuous support and genuine efforts to foster values of tolerance and equality in our family.

List of abbreviations

CERD	Convention on the Elimination of All Forms of Racial Discrimination
CEDAW	Convention on the Elimination of All Forms of Discrimination against Women
CoE	Council of Europe
DNS	Domain Name Systems
ECHR	European Convention on Human Rights and Fundamental Freedoms
ECtHR	European Court of Human Rights
ECRI	European Commission against Racism and Intolerance
EU	European Union
HRC	Human Rights Committee
IACtHR	Inter-American Court of Human Rights
ICCPR	International Covenant on Civil and Political Rights
ICANN	Internet Corporation for Assigned Names and Numbers
ISP	Internet Service Provider
ITU	International Telecommunications Union
SPLC	Southern Poverty Law Centre
UDHR	Universal Declaration of Human Rights
UN	United Nations
UNGPs	United Nations Guiding Principles on Business and Human Rights
UNHRC	United Nations Human Rights Council
US	United States of America
WCIT	World Conference of International Telecommunications

CHAPTER ONE

Introduction

1.1. Statement of the problem

"In the 1970s and 80s the average white supremacist was isolated, shaking his fist at the sky in his front room. The net changed that"[1]

- Mark Potok, a former editor at the SPLC

The Southern Poverty Law Centre (SPLC) - the primary organisation in the United States (US) that monitors and exposes hate groups and extremist activities both online and offline, recorded 954 active hate groups operating within the US territory in 2017[2]. Such groups are hardly alone in achieving their objective – they operate alongside like-minded groups across the globe as well as individuals who merely express their views, without having a specific goal of propagating hate. Had those individuals existed in a world without internet and more importantly, social media, the effects of their activities could have been less detrimental compared to the contemporary world. However, due to the transformation of the communications system, specific challenges arise that requires a re-evaluation of the existing strategies and standards.

The internet revolution has redefined the concept of communications. Alongside positive changes, such as, empowering individuals who were silenced before, the internet created an access to easy, cost-effective, quick communication irrespective or individuals' location[3]. Furthermore, the Internet created space for freedom of expression for everyone irrespective of

[1] Fear and loathing, available at: https://www.theguardian.com/uk/2004/aug/12/race.world (last visited 22 May 2018).

[2] The Southern Poverty Law Centre (SPLC), Hate Map, available at: https://www.splcenter.org/hate-map (last visited 22 May 2018)

[3] Balkin, J. M. (2004). Digital speech and democratic culture: A theory of freedom of expression for the information society. *NyuL rev.*, 79, 1. P. 1.

their status and capabilities even in the states that were reluctant to provide such opportunities in real life.

The area of communications further changed as social media platforms emerged in the 1990s[4]. Unlike the Internet, that provided an opportunity to certain groups of individuals to disseminate their ideas, social media platforms created equal opportunities for every single person in the world, who had access to the internet. Social media platforms simply became intermediaries between two groups of people - private users who share the content and the audience, who receives them. The means of communications became accessible to a great number of people.

However, this low-cost and high-speed dissemination mechanism had its drawbacks – it soon became an ideal venue to facilitate spreading hate speech. Soon after social media platforms became popular, a growing number of groups emerged, that was devoted to homophobic, Islamophobic, anti-immigrant, anti-Semitic hate, misogyny, white supremacy etc.[5] Social media provided an opportunity for radical groups to find like-minded individuals, to create collective identity and solidarity for a certain ideological viewpoint, To further connect with each other and collaborate. Consequently, the ability to influence the world through social media platforms increased and, as it has been argued, today's computer keyboards may be "even more destructive, than tanks and machine guns"[6].

Addressing hate speech on social media is a difficult task. It involves three sets of expression right that need to be considered while imposing constraints on social media: rights of the individuals who express the opinion, rights of the social media platforms and the third party

[4] Banks, J. (2010). Regulating Hate Speech Online. *International Review Of Law, Computers & Technology*, (3), 233.

[5] See, for example, Hanzelka, J., & Schmidt, I. (2017). Dynamics of cyber hate in social media: A comparative analysis of Anti-Muslim movements in the Czech Republic and Germany. *International Journal of Cyber Criminology*, 11(1) p. 147.

[6] Shaw, L. (2011). Hate Speech in Cyberspace: Bitterness without Boundaries. *Notre Dame Journal of Law, Ethics and Public Policy*, 25(1), 279–304, p. 279

readers. Furthermore, the right to equality of those who are victims of this kind of speech also needs to be taken into account.

To find the right balance between conflicting interests, it is crucial to analyse key concepts. First, the definition of hate speech is contested and may include a wide spectre of expression. Second, determining the most favourable form of social media governance is challenging.

The models that are suggested for the intervention in online hate speech mainly focus on social media platforms. They have a unique role as they are intermediaries providing individuals with access to the digital sphere[7]. However, states frequently intervene through legislative or non-legislative methods to pressure social media platforms.

1.2. Why does online hate speech matter

Recent work has highlighted the repercussions of online hate. The relationship between hate speech and violence has been evidenced in history. Hate speech was a major tool employed to promote slavery in Colonial America, to aggravate tensions in Bosnia and in the rise of the Third Reich[8]. The aim of such speech is to ridicule victims, to humiliate them and represent their grievances as less serious[9]. The UN HRC Special Rapporteur on Minority issues states that even if hateful messages do not materialize in an actual crime, they create a precondition for a hate crime as hate crimes are most likely to occur with prior stigmatization and dehumanization of targeted victims[9].

The main question related to online hate speech is whether its effects can be traced to real life events. A considerable amount of literature has been published on the issue and the findings

[7] DeNardis, L., & Hackl, A. M. (2015). Internet Governance by Social Media Platforms. *Telecommunications Policy*, 39(9), 761-770.

[8] Shaw, L. (2011). Hate Speech in Cyberspace: Bitterness without Boundaries. *Notre Dame Journal of Law, Ethics and Public Policy*, 25(1), 279-304. [9] *Ibid.*

[9] Report of the Special Rapporteur on minority issues, Rita Izsák - Hate speech and incitement to hatred against minorities in the media, A/HRC/28/64, 5 January 2015

are rather disappointing - the possible conclusion could be that online hate speech hardly ever stays purely virtual[10]. Hateful speech, even if it does not reach the threshold of "incitement to violence", can be detrimental and reinforce the negative, biased beliefs in the society[11]. Not only does it intensify prejudice and stereotypes but also affects the mental health of the targeted individuals. Different studies point out that negative feelings towards minorities and stereotypes tend to increase with time and it only takes a "trigger" event to result in hate crimes. An example of such speech could be events that took place in response to the 9/11 attacks in the US - previous anti-Muslim rhetoric turned out to be detrimental and result in hate crimes[12]. In this case, prejudicial motives materialized aftermath this event and the previous attempts to represent Muslims as terrorists were used as a basis to create a hostile environment against them.

Legal philosopher Jeremy Waldron identifies two dangerous types of messages in hate speech that exposes different groups to vulnerability[13]. The first message is directed at the victims and intends to dehumanize or ridicule them and to make them feel unwelcome in the society[14]. Similarly, the overall effect of hate speech is to insult victims, stereotyping them, for example, as terrorists, advocating the exclusion of them from society, denying them human rights, holding them accountable for the actions of the other members of the group, applying double standards etc[15].

[10] Jakubowicz, A., Dunn, K., Mason, G., Paradies, Y., Bliuc, A. M., Bahfen, N., ... & Connelly, K. (2017). Cyber Racism and Community Resilience: Strategies for Combating Online Race Hate. Springer, p. 41.
[11] Shaw, 2011, p. 279.
[12] Williams, M. L., & Burnap, P. (2016). Cyberhate on Social Media in the Aftermath of Woolwich: a Case Study in Computational Criminology and Big Data. *British Journal of Criminology*, 56(2), 211-238, p. 214.
[13] Waldron, J. (2012). The harm in hate speech. Cambridge, Mass.; *London: Harvard University Press*, 2012, pp. 2-3.
[14] *Ibid.*
[15] Jakubowicz, A., Dunn, K., Mason, G., Paradies, Y., Bliuc, A. M., Bahfen, N., ... & Connelly, K. (2017). Cyber Racism and Community Resilience: Strategies for Combating Online Race Hate. Springer, p.42.

The second message, on the other hand, is aimed at the rest of society and intends to encourage people into thinking that there are some like-minded individuals who agree with the idea that the certain groups of society should be excluded and not tolerated[16].

Moreover, hate speech, once it appears online, does not disappear afterward and opens a possibility to be present and reused on social media for unlimited period of time. As Andre Oboler, from the Online Hate Prevention Institute, has noted,

> "The longer the content stays available, the more damage it can inflict on the victims and empower the perpetrators. If you remove the content at an early stage you can limit the exposure. This is just like cleaning litter, it doesn't stop people from littering but if you do not take care of the problem it just piles up and further exacerbates [it]"[17].

Studies of the effects of online hate speech show that the greater danger, nevertheless, can stem from the normalisation of hate through social media[18]. The aim of the hate groups is not only to publish the content that contains messages but also to make such content appear as a normal part of society[20]. If the hateful message can be perceived by the society as just another opinion on social media then such hate can be openly expressed. The overall objective is to create "social acceptability" regarding the hate content.

1.3. Objective and research question

The main objective of this thesis is to explore the role of social media platforms in promoting intolerance, to examine the existing legal remedies to address the problem and to identify the

[16] *Ibid.*
[17] Online Hate Prevention Institute. 2014. "Press release: Launch of online tool to combat hate". Available at: http://ohpi.org.au/press-release-launch-of-online-tool-to-combat-hate/ (Last accessed 22 May 2018).
[18] Jakubowicz, A., Dunn, K., Mason, G., Paradies, Y., Bliuc, A. M., Bahfen, N., ... & Connelly, K. (2017). Cyber Racism and Community Resilience: Strategies for Combating Online Race Hate. Springer. p. 43 [20] *Ibid.*, p. 45.

gaps in the social media governance that leaves a significant part of the society vulnerable to hate propaganda.

Examining the role of social media regulation in relation to online hate speech is crucial. The thesis addresses the dichotomy between "governance of social media" and "governance by social media" and therefore provides an analysis of flaws in the existing model. The thesis will explore how national mechanisms or international legal instruments attempt to regulate social media and how social media platform policies constitute privatized governance. Further focus will be on a critical assessment of the protection of human rights in the digital environment. The research questions of this thesis are the following:

1. What are the major obstacles in regulating hate speech on social media?
2. What are the implications of the protection of individual rights by private intermediaries rather than by governments or global institutions?

1.4. Methodology

The core focus of the thesis is analysing the implications of private regulation on hate speech on social media. To achieve this aim, this paper will employ various legal research methods. With regards to identifying the definition of hate speech, doctrinal, comparative and critical methods will be used to establish the different approaches suggested by different international and regional human rights treaties. Thus, the thesis will have a core focus on the International Covenant on Civil and Political Rights (ICCPR) and International Convention on the Elimination of All Forms of Racial Discrimination (CERD). With regards to regional documents, the European Convention on Human Rights (ECHR) will be examined. Alongside the human rights treaties, the thesis will also include the documents and definitions provided by the relevant jurisprudence and interpretative bodies.

Chapter 3 of this thesis provides an overview of classical and contemporary theoretical frameworks regarding freedom of expression. This section will mainly employ dogmatic

research methodology, focusing on the evaluation and analysing the work of free speech scholars, including Mill, Mackinnon, Dworkin, Balkin, etc.

The nature of hate speech intrinsically requires inter-disciplinary research methods. Especially, the comprehensive analysis of hate speech on social media and its effects could hardly be conducted only through doctrinal methodology. This type of analysis will particularly be beneficial to understand the implications of hate speech that can be disseminated very quickly without national border restrictions and by anyone, to further understand theoretical challenges imposed by hate speech. Therefore, the socio-legal method will be employed to analyse the works of communication, social and media studies. In that respect, the works of the academics, such as Cohen-Almagor, DeNardis, Klein, Oboler, and Waldron will be analysed.

Throughout the thesis, certain sections will further employ the case-study method to demonstrate the difficulties arising regarding interpretation and application of legal rules of hate speech on social media.

1.5. Definition of key terms

It needs to be acknowledged that finding a holistic definition of *"hate speech"* is a challenging task as the term itself is vague and subject to various interpretations. As it will be illustrated below, the definition of hate speech largely depends on a context and a jurisdiction, accordingly, coming up with a clear answer is a difficult task. However, it is still possible to place boundaries and identify definition. The thesis will rely on the broad classification suggested by The Council of Europe (CoE) Committee of Ministers, that considers "all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance, including: intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility against minorities, migrants and

people of immigrant origin"[19]. The main reason for this choice it is twofold: first, the definition is provided by an authoritative regional body and accordingly, has a significant legal weight and second, it leaves an opportunity for relatively wide-ranging discussions of what is and should be considered as hate speech, among others, by social media platforms.

Throughout this paper, *"social media"* will refer to "a group of Internet-based applications built on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of User Generated Content[20]" "Social media" is an umbrella term and can take many forms, including blogs (e.g., Twitter), content communities (e.g., YouTube), social networking sites (e.g., Facebook), collaborative projects (e.g., Wikipedia) etc[21]. Among the different social media platforms, Facebook and Twitter are among the most popular with billions of users[22]. Online collaboration platforms, such as Google Docs can also be considered as a form of social media if they enable a number of people to cooperate and convey a message similar to an event that took place during Egyptian uprising in 2011[23]. Similarly, emails and private text messages are excluded from this definition as they are not publicly available, however, mass texting or emailing can under certain circumstances be considered similar to social networking sites as they result in immediate dissemination of information to a large audience.

"Web 2.0" refers to Internet platforms that allow for interactive participation by users, while *"user-generated content*" is the name for all of the ways in which people may use social media[24].

"*Internet governance*" is a broader term frequently used to refer to the design and administration of the technical infrastructure that is necessary to ensure sufficient functioning

[19] Council of Europe, Recommendation No. R 97 (20) of the Committee of Ministers to Member States on "hate speech", adopted on 30 October 1997.
[20] Sangsuvan, K. (2014). Balancing Freedom of Speech on the Internet under International Law. *North Carolina Journal of International Law & Commercial Regulation*, 39701, p. 721.
[21] *Ibid*. p. 723.
[22] Joseph, S. (2012). Social media, political change, and human rights. BC Int'l & Comp. L. Rev., 35, p. 147.
[23] *Ibid*.
[24] *Ibid*.

of the Internet[25]. The term "*social media governance*" is a relatively narrower concept that is related to platform design choices and user policies[26]. In the thesis, these terms might be used interchangeably only if the broad concept of Internet regulation includes applies to social media governance as well.

1.6. Thesis structure

The thesis consists of seven main chapters. *Chapter 1*, the introduction is followed by unraveling the term "hate speech" and consider its definitions in the international and regional human rights treaty law. *Chapter 2* describes general principles of freedom of expression in international human rights law, including the UDHR, the ICCPR and the ECHR etc. Further, this chapter examines whether principles regarding hate speech apply to the expression through social media platforms.

Chapter 3 examines different theoretical frameworks adopted in relation to freedom of expression and both its classic and contemporary understanding. This chapter aims to analyse how different legal theories help to approach freedom of expression. This chapter introduces relatively new theoretical approaches towards freedom of expression in the digital age and suggests a reassessment of the existing theoretical framework.

Chapter 4 provides a rather descriptive analysis of the problem, that is, what are the implications of hate speech through social media platforms. It reflects on the existing challenges hate speech poses on social media, the contemporary forms through which hate speech is materialized and the consequential negative impact on its victims, who at the same time predominantly belong to minority groups. It is followed by an explanation of the process of dissemination of expression through social media platforms.

[25] DeNardis, supra note 7, p. 1.
[26] DeNardis, supra note 7, p. 2.

Chapter 5 explores whether and how hate speech on social media is addressed to protect members of the society. The thesis looks at the different definitions of hate speech adopted by social media platforms and provide criticism of the practical application of those rules. The question with regards to strategies adopted against hate speech on social media is twofold. First, the problems arising from private regulation are examined, and then contrasted with the implications resulting from government formal and informal pressures of social media platforms.

Next, *Chapter 6* of the thesis looks at the role of the international and regional bodies. This chapter examines the possibility of regulation on the international level, past attempts to do so and the major obstacles that have hindered moving to more cooperative mechanisms.

The last part, *Chapter 7*, summarises and concludes the main findings of the thesis and provides final remarks for the future consideration.

1.7. Delimitations

For the purposes of this thesis, it is essential to distinguish between legal, moral and social responses to hate speech on social media. Legal responses refer to addressing issues through institutions. Moral and social responses, on the other hand, are related to societal implications of a given conduct, such as taking responsibility by individuals or social media platforms to refrain from acting when it may harm the society. In other words, it means taking responsibility to have a more advanced society. Much of the literature focuses on non-legal measures to reduce the effects of hate speech, such as counter-speech or imposing social liability on key social media platforms, introducing more cooperative system etc. However, the core focus of the thesis will be legal measures and strategies adopted against hate speech on social media, without questioning the necessity of non-legal measures.

The issue of hate speech on social media can generally be analysed from different perspectives. The questions regarding this issue can be classified into different categories, such as “what should be regulated”, “who should regulate” and “according to which rules”. Although the

main focus of this thesis will be to discuss and analyse the latter two questions, with regards to the first question – the content of hate speech that needs to be regulated, will also be addressed. Nevertheless, in that respect, the paper will be relatively brief and rely on the existing international human rights standards.

The possibility of abusing hate speech regulations by oppressive states also needs to be acknowledged. Hate speech restrictions can be applied by governments to silence the voices of minority or worse, to promote majority oppression in the name of defending someone's dignity. However, this thesis will mainly analyse the problem of hate speech in social media in the context of states that are in favour of human rights and democratic values. The main rationale behind it is to demonstrate that the existing model of regulation of hate speech is problematic even when the states promote human rights.

Furthermore, throughout this paper the term "hate speech" might encompass the expression that does not primarily disseminate hatred, but serves such aim. Defamation and "fake news" are the examples that can contribute to representing the image of the minorities as, for example, violent and threatening. Although defamation and "fake news" are topics that independently be explored, as a different topic, it is unavoidable to connect them to the problem of hate speech, especially in the age of social media.

This paper does not intend to question the issue of the liability of social media platforms, in other words, intermediary liability for the content that is produced and disseminated by a third party. Although the concept of the intermediary liability could be fairly arguable, this thesis relies on the jurisprudence of the ECtHR, according to which, although private intermediaries simply provide a platform for an individual expression, they still remain responsible for the expression [27] . Furthermore, imposing a liability for third-party content has been a

[27] See, for example, Delfi AS v. Estonia (GC), no. 64569/09, 16 June 2015 (The case concerned comments and anti-Semitic threats in the comment section of the Estonian Journal, named Delfi. The ECtHR found the online journal responsible for those comments, even though the Journal had a solid record of removing such comments).

wellestablished practice with regards to media and publishing organisations[28]. Therefore, no further analysis will be provided from that perspective.

Regarding the overview of the regional human rights standards, the core focus of the thesis will be the European system for a number of factors, but among them, the primary reason to do so is due to its historical context. However, guarantees of freedom of expression provided by other regional documents will also be covered.

Furthermore, the thesis does not aim to provide conclusive answers to the questions of how should hate speech regulated on social media and who has the best capabilities to adopt strategies against it. Finding a solution would be a rather difficult task and largely dependent on the willingness of sovereign states to cooperate. Rather, the thesis will focus on identifying and demonstrating the key legal problems related to the regulation and focus on the necessity to revisit the existing approaches.

The global and cross-border nature of hate speech on social media needs to be acknowledged. However, its effects can vary based on the context they are disseminated in. Due to the limitations of time and resources, it would be impossible to provide a comprehensive research of the forms and types of expression worldwide. Therefore, the thesis relies on the particular examples from specific countries with the aim to cover different areas of the world. Such approach always bears a risk of making generalised statements. However, exploring the research question otherwise would be difficult.

[28] See, for example, case of N.Y. Times Co. v. Sullivan, 376 U.S. 254 (1964)

CHAPTER TWO

Definition of "hate speech"

The critical factor to further examine the problem of hate speech on social media is to determine what can be considered as hate speech to start with. While the particular examples of hate speech will be considered below, this chapter aims to examine international human rights law and identify the approach adopted so far. The question is twofold. First, it is necessary to analyse the definition of hate speech under international human rights treaties and their interpretive bodies. Second, it is essential to identify whether these rules apply to the online and offline world equally. This chapter will provide the analysis of the relevant framework under international human rights treaties.

2.1. International human rights framework

In the post-World War II period, when a number of international instruments were adopted, the necessity of recognizing the right to free expression was acknowledged[29]. Freedom of expression is considered as one of the major human rights, thus recognized by every relevant international human rights treaty. Freedom of expression is not an absolute right and the grounds of limitations are also enshrined in the documents[30]. However, with regards to online hate speech human rights treaties are silent, as they primarily emerged from the period when the transformation of the sphere of communications by social media and its effects on human rights were unforeseeable[31]. Therefore, the early international human rights instruments, such as the UDHR and the ICCPR do not address the issue of online "hate speech".

[29] See, for example, ICCPR, UDHR.
[30] See, for example, Article 19(3) of the ICCPR).
[31] Coe, P. (2015). Social Media Paradox: An Intersection with Freedom of Expression and the Criminal Law, *Information & Communications Technology Law*, (1), 16, p. 18.

To compound the problem, the primary human rights documents rarely provide any specific reference to "regular" hate speech, meaning hate speech in the offline world, or any further definition of what could be considered as such. The 1948 Universal Declaration of Human Rights (UDHR) exemplifies the issue. The UDHR the first non-binding document adopted right after the end of Holocaust was intended to promote the protection of human rights to leave behind the violations in the past[32]. The UDHR attempted to find the right balance between the rights to equal treatment and freedom of expression. On the one hand, the UDHR contains the values of fundamental importance and recognises the rights to equal protection under the law which indicates that:

> "All are entitled to equal protection against any discrimination in violation of this Declaration and against any incitement to such discrimination[35]".

However, the UDHR also proclaims that everyone has the right to freedom of expression, which includes:

> "Freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media and regardless of frontiers"[33].

The UDHR does provide the grounds for limiting freedom of expression. On the other hand, hate speech is not considered as such ground. The text does not explicitly refer to hate speech. The same attitude was adopted by the ICCPR, that is claimed to be one of the most important and comprehensive international instruments when addressing hate speech[34]. Although it does not explicitly mention the term "hate speech", the document specifically addresses two different kinds of hate speech. It contains twofold safeguards - by protection to the right to

[32] Sangsuvan, K. (2014). Balancing Freedom of Speech on the Internet under International Law. North Carolina Journal of International Law & Commercial Regulation, 39701, p.709. [35] UDHR, Art 7.
[33] UDHR, Art 19
[34] See, for example, Sangsuvan, K. (2014), p. 709.

freedom of expression in Article 19 and the prohibition of advocacy to the hatred that constitutes incitement to discrimination, hostility or violence in Article 20[35].

The International Convention on the Elimination of all Forms of Racial Discrimination, adopted in 1965, was the first international treaty explicitly dealing with the issue of hate speech[36]. Article 4 contains a very broad obligation and requires states to:

> "condemn all propaganda and all organisations which are based on ideas of theories of superiority of one race or group of persons of one colour or ethnic origin, or which attempt to justify or promote racial hatred and discrimination in any form".

In the General Comment No. 15, the CERD Committee distinguished four different forms of hate speech and among the most radical forms, it included the dissemination of ideas based on racial superiority or racial hatred as the prohibited form of hate speech according to Article 4(a)[37]. The CERD definition of hate speech is the most far-reaching definition so far, however, it is limited to a certain ground of hatred and does not cover hate speech on other grounds, such as gender, sexual orientation etc[41]. The document itself relates to racial discrimination and the application of the principles to other forms of discrimination can be arguable. The CERD committee has further recognised hate speech as an expression of hate on the grounds other than Article 1, such as religion, gender etc., although on a treaty level, there is no such protection of those minorities. [38]

Other human rights treaties, for example, the 1979 Convention on the Elimination of all forms of Discrimination Against Women (CEDAW) does not explicitly require prohibitions on "hate speech" against women, however, imposes obligation on states to combat discrimination to eliminate prejudices and all other practices "which are based on the idea of the inferiority or

[35] Article 19(3) and 20 of the ICCPR.
[36] General Assembly Resolution 2106A (XX), 21 December 1965, entered into force 4 January 1969.
[37] General Comment No. 15 of 23 March 1993, on article 4 of the Convention, the CERD Committee, para. 3.
[41] Article 1 of the CERD defines race as a broad term and includes other characteristics, such as color, descent, or national or ethnic origin.
[38] The CERD Committee, Concluding Observations on Romania, CERD/C/ROU/CO/16-19, 13, September 2010, para 4.

the superiority of either of the sexes or on stereotyped roles for men and women"[39]. Nevertheless, the provision is vague and leaves the actions that might cause the consequences of discrimination to an interpretation.

The abovementioned treaties, as illustrated, adopt different terms regarding the state's obligations in relation to combating hateful expression. These treaties do not provide exhaustive characteristics of speech that should be prohibited. Therefore, the limitations on freedom of speech can be defined by analysing different treaties. Some of these terms are rather permissive while others limit the discretion of a state and impose direct obligations to restrict speech.

The UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression in his 2012 Annual Report provides a classification of hate speech and distinguishes between three types of expression: expression that needs to be prohibited as it constitutes an offence under international law; expression that may be prohibited, even though it does not constitute a criminally punishable act; and expression that constitutes hate speech and raises concerns in terms of tolerance, but is still justified under international law[40]. The following section will analyse each of them and further elaborate on what kind of speech is prohibited under international law.

2.1.1. Restricted hate speech

The international human rights law is less ambiguous with regards to what constitutes the restricted hate speech. It acknowledges the effects of the most severe forms of hate speech and thus obliges states to prohibit these types of expression. A clear example of such prohibition under international law is the direct and public incitement to genocide - an act prohibited by the 1948 Convention on the Prevention and Punishment of the Crime of Genocide, alongside

[39] Article 5 of the CEDAW.
[40] The annual report of the UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression (Special Rapporteur on FOE) to the General Assembly, (2012), A/76/357, para. 2.

with the prohibition of genocide itself[41]. However, the additional preconditions are the public nature of the statements and the direct communication.

Hate speech that must be prohibited is further given in Article 20 of the ICCPR. However, as the ICCPR provides two articles specifically dealing with the limitations on freedom of expression, namely, Article 19 and Article 20, first it is necessary to examine their relationship.

Article 19 of the ICCPR recognizes the right to freedom of expression and sets down limitations of this right. According to Article 19(3), freedom of expression is subject to restrictions to protect the rights or reputations of others, national security or of public order (*ordre public*), or of public health or morals. Article 20 further explicitly limits freedom of expression in cases of

> "advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence"[42].

The annotations to the 1955 draft of the ICCPR show that this article and its particular aspects regarding hate speech has been deeply contested. Some states argued that the general limitation clause given in Article 19 paragraph 3 was sufficient enough to deal with hate speech, while others considered a separate provision (Article 20) necessary to expressly deal with hatred that constitutes incitement to harm[43]. The main argument against Article 20 was that the limitations on freedom of expression could be abused by some states[44]. Even after the

[41] The Convention on the Prevention and Punishment of the Crime of Genocide, 9 December 1948.
[42] *Ibid*.
[43] Annotations on the text of the draft International Covenants on Human Rights, A/2929, 1 July 1955, paras. 189-194. Available at: http://www2.ohchr.org/english/issues/opinion/articles1920_iccpr/docs/A-2929.pdf (Last visited 22 May 2018).
[44] *Ibid*.

document was finalised, some states, including Australia, New Zealand, the United Kingdom and the United States, refused to ratify it and attached reservations to Article 20[45].

The difference between Article 19 (3) and Article 20, aside from the different protections provided, is that the limitations to freedom of expression under the first one are optional, while the second imposes obligatory limitations. Article 19(3) states that freedom of expression

"may, therefore, be subject to certain restrictions", while Article 20 underlines that any advocacy of hatred that constitutes incitement to discrimination, hostility or violence "shall be prohibited by law"[46].

Even though there is a clear tension between these two articles, the UN Human Rights Committee (HRC) has specifically addressed the conflict and stated that Article 19(3) and Article 20 are compatible[47]. As a result, the Article 20(2) can only be applied if it meets the requirements under Article 19(3). Therefore, Article 20 of the ICCPR provides an absolute prohibition of any advocacy of discriminatory hatred that constitutes incitement to discrimination, hostility or violence.

The UN Special Rapporteur further provided a definition of what could be considered as a hate speech under Article 20 of the ICCPR. Advocacy of hatred on the bases of different grounds does not constitute an offense *per se.* Such advocacy becomes an offense only when the preconditions under Article 20 are present, such as incitement to discrimination, hostility or violence, or "when the speaker seeks to provoke reactions on the part of the audience"[48].

[45] Status of treaties, available at: https://treaties.un.org/Pages/ViewDetails.aspx?src=IND&mtdsg_no=IV4&chapter=4&clang=_en (Last visited 22 May 2018).
[46] Article 19 and 20 of the ICCPR.
[47] General Comment 11: Prohibition of propaganda for war and inciting national, racial or religious hatred (Art. 20), 29 July 1983.
[48] The annual report of the UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression (Special Rapporteur on FOE) to the General Assembly, (2012), A/76/357, para.43.

The Special Rapporteur further explained that the term "incitement" includes an element of imminence, i.e. risk of discrimination, hostility or violence need to be real and present[49]. Consequently, this article covers only special, narrow scope of expression and would not extend to any kind of hateful expression. States enjoy a wide margin of independence to decide whether a particular expression constitutes hate speech.

The Rabat Plan of Action was later adopted in 2012 by experts following a series of consultations adopted by the OHCHR and contains recommendations for the implementation of Article 20(2) ICCPR[50]. The document contains a six-part threshold test to define circumstances under Article 20(2), namely, the context of the expression, identity of the speaker, intent of the speaker to advocate hatred, the content of the expression, extent, and magnitude of the expression and likelihood of imminent harm occurring.

2.1.2. Hate speech that may be restricted

While Article 20 requires a very high threshold, relatively lower restrictions are provided under Article 19 of the ICCPR. Article 19 provides the three-part test for restrictions and includes legality, proportionality and necessity requirements. Accordingly, these restrictions must be provided by law, pursue a legitimate aim, and be necessary for a democratic society[51].

ARTICLE 19, a British human rights organisation, further explains that the threshold provided by Article 20(2) is too stringent and the requirement of "incitement" is only present in limited circumstances[52]. That's when different grounds enlisted in Article 19(3) comes into the play,

[49] *Ibid*, para. 44.
[50] The Rabat Plan of Action on the prohibition of advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence, A/HRC/22/17/Add.4, Appendix, adopted 5 October 2012.
[51] Article 19 of the ICCPR.
[52] ARTICLE 19, (2015) 'Hate Speech' Explained: A Toolkit. Available at: https://www.article19.org/data/files/medialibrary/38231/'Hate-Speech'-Explained---A-Toolkit-%282015Edition%29.pdf (Last visited 22 May 2018), p.84.

in particular, when it concerns the "rights of others" and "public order"[53]. Such hate speech may target both – groups and individuals.

The case of *Malcolm Ross v. Canada* is a landmark case of the UN Human Rights Committee, that justified limiting freedom of expression on the ground of the rights of others[54]. The case concerned a teacher in school who published controversial books denying Holocaust and Jewish religion. Due to the threat of inciting discrimination and creating "poisoned school environment", the teacher was removed from the office[55]. The UN HRC found the removal justified under Article 19(3) of the ICCPR as it was proscribed by the legislation, served a legitimate aim and was necessary "to protect the right and freedom of Jewish children to have a school system free from bias, prejudice and intolerance"[56].

2.1.3. Lawful hate speech

An expression that does not meet the requirements provided under Article 19(3) can be considered to be a lawful form of hate speech. The UN Human Rights Committee has affirmed that under Article 19 of the ICCPR, the expression of opinions and ideas are protected, even if they might be received by someone as deeply offensive[57]. Therefore, this right may encompass discriminatory expression.

The Special Rapporteur further emphasized that the limitations recognized under Article 19 are not intended to "suppress the expression of critical views, controversial opinions or politically incorrect statements"[58]. Broad interpretations of those limitations would not

[53] *Ibid*.
[54] Malcolm Ross v. Canada, CCPR/C/70/D/736/1997, UN Human Rights Committee (HRC), 26 October 2000.
[55] *Ibid*.
[56] *Ibid*., para. 11.6.
[57] See HR Committee, General Comment No. 34 on Article 19: Freedoms of opinion and expression, CCPR/C/GC/34, 12 September 2011, para 11.
[58] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Promotion and protection of all human rights, civil, political, economic, social and cultural rights, including the right to development. A/HRC/36/1, para. 24.

comply with the international instruments and threaten the full enjoyment of rights[59]. Otherwise, measures that are not clearly and narrowly defined, "there is a very real possibility of such measures being abused, particularly where respect for human rights and democracy is weak, and "hate speech" laws have in the past been used against those they should be protecting"[60].

For example, in *A.W.R.A.P v. Denmark*, the CERD Committee found no violation on cases that concerned statements directed to religion rather than individuals[65]. The statements aimed to criticize Koran and Islam in general, attacking the entire culture. On the other hand, statements of the petitioners did not target specific groups or individuals. Although the Committee underlined the negative nature of the comments no violation of the Article 4 of the CERD was founded.

Accordingly, when hate speech falls within the scope of lawful expression, it does not automatically mean that states do not have an opportunity to adopt non-legal measures and policies to address this type of hate speech. States still need to adopt some measures to combat negative stereotypes, other than restricting expression. Such requirements are enshrined in different legal documents[61].

[59] *Ibid*., para. 66.
[60] Joint statement by the United Nations Special Rapporteur on freedom of opinion and expression, the OSCE Representative on freedom of the media and the OAS Special Rapporteur on freedom of expression on Racism and the Media, Available at: https://www.osce.org/fom/40053?download=true (Last visited 22 May 2018). [65] A.W.R.A.P v. Denmark, The Committee on the Elimination of Racial Discrimination, no. 37/2006, 8 August 2007.
[61] See, for example, Article 5 of the CEDAW.

2.2. The regional human rights framework

Regional human rights treaties—the European Convention on Human Rights (ECHR)[62], the American Convention on Human Rights (ACHR)[63], and the African Charter on Human and Peoples' Rights (ACHPR)[64]—guarantee the right to freedom of expression similar to the ICCPR. Among them, the ACHR is the only document specifically referring to the banning of hate-motivated speech. According to Article 13(5), advocacy of national, racial or religious hatred that constitutes incitement to lawless violence need to be banned.

The major focus of this section will be the jurisprudence of the European Court of Human Rights (ECtHR). Under Article 10(1) of the ECHR, the right to freedom of expression is guaranteed with a few exceptions. However, under Article 10(2), the grounds to limit freedom of expression does not specifically include the examples of hate speech, rather it allows prohibition of expression on more general grounds, such as the protection of the rights of others. Therefore it is possible to conclude that the ECHR does not impose any obligations on states to limit hate speech. States are granted a certain level of discretion to adopt appropriate regulations and those regulations will further be subject to supervision by the Court[65].

However, the Court underlined that certain forms of speech should not be protected to meet the objectives of the Convention as a whole[66].

With regards to the practical application of the recognized principles, the court has a wellestablished case law. Interestingly, the term - hate speech was used by the ECtHR in 1999 for the first time, but the Court was reluctant to provide an explanation of hate speech[67]. Since then, the Court has found that the term "hate speech" has an autonomous meaning and that the

[62] Article 10 of the ECHR, (1950).
[63] Article 9 of the ACHR, (1981).
[64] Article 13 of the ACHPR (1981).
[65] *Ibid*, p.8.
[66] See: Erbakanv . Turkey, App. No. 59405/00 (2006), para. 56.
[67] Sürek v. Turkey (No. 1), para. 62.

analysis of each case needs to be based on its own merits, therefore, the definitions could impose limitations in future cases[68].

Analysis of the established case-law of the European Court of Human Rights on hate speech shows that the court recognizes a relatively low level of protection to speech that involves hatred. However, the court has applied two different standards to different forms of hate speech, namely, the speech that has different targets. In some cases, the Court finds that the speech does not fall under the protected sphere of Article 10 and is inadmissible under the prohibition of abuse of rights clause, while in others, the speech simply contradicts Article 10 and finds a violation of that Article.

An example of the application of different standards could be the case *Pavel Ivanov V Russia* where the applicant published articles in his own newspaper blaming Jews for being the source of all evil in Russia and plotting a conspiracy against the Russian people[69]. As a result, he was convicted on the ground that he has conducted public incitement of hatred. The Court declared the application inadmissible as underlying that attack on Jews did not constitute "speech" therefore it did not fall within the protected sphere of Article 10. Moreover, the Court emphasized that such verbal attack against Jews was against the core values of the Convention, such as tolerance, social peace, and non-discrimination[75]. The speech of the applicant, therefore, contradicted Article 17 of the Convention that prevents the use of the Convention rights to 'engage in any activity or perform any act aimed at the destruction of any of the rights and freedoms' in the Convention.

Article 17, ECHR provides prohibition of abuse of rights clause. The Court generally applies this clause to guarantee that Article 10 protection does not extend to racist, xenophobic or anti-

[68] McGonagle, T. (2013). The Council of Europe against online hate speech: Conundrums and challenges. Council of Europe Conference Expert Paper, p. 35.
[69] Pavel Ivanov v Russia, Application no. 35222/04 (2007).
[75] *Ibid*.

Semitic speech, statements that deny, dispute or minimise the Holocaust[7071]. The Court consistently declares cases regarding this type of expression as unfounded and inadmissible.

The Court's case-law, nevertheless, is not consistent and in particular cases, it has adopted a different approach with regards to speech containing homophobic hate and ethnic hatred[77]. In such cases, the ECHR instead considers speech as an infringement under Article 10 and therefore, the three-stage test under Article 10(2) is applied.

Unlike the ECHR, the Council of Europe (CoE) Committee of Ministers has adopted a broader approach and defined hate speech as follows:

> "the term "hate speech" shall be understood as covering all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, anti- Semitism or other forms of hatred based on intolerance, including: intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility against minorities, migrants and people of immigrant origin"[72].

This provision implies that the classical "incitement" requirement that is a crucial precondition to identify hate speech under other major human rights treaties is only an alternative requirement under the CoE Recommendation. Other forms of expression, such as spreading, promoting and justifying hatred would constitute hate speech. Therefore, the protection provided by the European level is much higher and covers other consequences of hate speech rather than only inciting to violence suggested by other international human rights treaties.

[70] See, for example, Garaudy v. France, application.no 65831/01 (2003), Seurot v. France, application no.
[71] /00 (2004) [77]See, for example, Vejdeland and Others v Sweden application no. 1813/07 (2012), Balsytė - Lideikienė v. Lithuania Application no. 72596/01, (2009).
[72] Council of Europe, Recommendation No. R 97 (20) of the Committee of Ministers to Member States on "hate speech", adopted on 30 October 1997, appendix.

2.3. Summary of the provisions regarding hate speech

The overall conclusion that can be drawn from the analysis of international and regional human rights treaties is that universally accepted definition of hate speech and obligations does not exist. The questions of how to draw boundaries between restricted and permitted hate speech or what constitutes "incitement" are extremely complex ones and has been a concern of many public and academic figures[73]. Moreover, while the requirement of "incitement" is crucial under some treaties, it does not bear the same weight under others. For example, the CoE provides a much broader definition compared to other international legal instruments[74].

Furthermore, the application of the human rights norms differs depending on the groups that are targeted, as demonstrated by analysing the case-law of the ECtHR and the interpretatory documents of the CERD[75]. These particular cases exemplify the absence of uniformity even on the international and regional level.

The lack of clear definitions is partially understandable due to the nature of the human rights treaties. Providing a comprehensive description of each term and a concept could be impossible and would require a reasonable interpretation depending on the circumstances and context. The main concern regarding the definition of hate speech on the international level is that as illustrated above, the scope of protection provided by different treaties differs significantly.

Due to the lack of the universally accepted definition of hate speech, international and national bodies are enabled to simply adopt their own, subjective definitions[76]. Nowak has pointed out that the lack of uniformity, the extraordinary vagueness of the terms such as "incitement" and

[73] See, for example, Study of the United Nations High Commissioner for Human Rights compiling existing legislations and jurisprudence concerning defamation of and contempt for religions, UN Doc. A/HRC/9/25, 5 September 2008, para. 24

[74] Council of Europe, Recommendation No. R 97 (20) of the Committee of Ministers to Member States on "hate speech", adopted on 30 October 1997.

[75] See, for example, the CERD Committee, Concluding Observations on Romania, CERD/C/ROU/CO/16-19, 13, September 2010, para 4.

[76] CoE Factsheet 'Hate speech' (2012), available at: https://www.echr.coe.int/Documents/FS_Hate_speech_ENG.pdf (Last visited 22 May 2018)

"advocacy" constitutes the risk of abuse[77]. Further shortcomings of such uncertainty could be a selective application of certain rules and the chance of state arbitrariness. Consequently, understanding of hate speech in relation to the information disseminated through digital means is a difficult process and has been referred to as "a jurisdictional and human rights nightmare"[78].

2.4. Whether these principles apply to social media

Even though the content of "hate speech" is ambiguous and controversial, as demonstrated above, its applicability to the online world, including social media is less disputed. Article 19 (2) of the ICCPR guarantees the right to freedom of expression, which includes "freedom to seek, receive and impart information and ideas of all kinds, regardless of frontiers, either orally, in writing or in print, in the form of art, or through any other media of his choice"[85].

The reference to "any other media of his choice" makes it clear that freedom of expression extends to different forms of technologies. It has been further recognised by the UN Human Rights Committee that freedom of expression applies to new forms of technology, including the Internet[79]. The UN Human Rights Committee in its General Comment 34 specifically addresses to the Internet and other mobile technologies and urges States to "take all necessary steps to foster the independence of these new media and to ensure access of individuals thereto"[80].

Regarding online speech under Article 19 (3) of the ICCPR, the Human Rights Committee has further provided an information in General Comment 34, underlining that any restrictions on

[77] Nowak, M. (2005). U.N. Covenant on Civil and Political Rights, CCPR Commentary, Kehl, *NP Engel Publisher*, p. 472.
[78] Siegel, M. L. (1998). Hate Speech, Civil Rights, and the Internet: The Jurisdictional and Human Rights Nightmare. *Albany Law Journal Of Science & Technology*, (2), 375 [85] Article 19 (2) of the ICCPR.
[79] Human Rights Committee General Comment 34, para. 12.
[80] *Ibid*., para. 43.

Internet-based platforms, including social media, are only permissible to the extent that they are compatible with paragraph 3. According to the General Comment,

> "Permissible restrictions generally should be content-specific; generic bans on the operation of certain sites and systems are not compatible with paragraph 3. It is also inconsistent with paragraph 3 to prohibit a site or an information dissemination system from publishing material solely on the basis that it may be critical of the government or the political social system espoused by the government"[81].

Similarly, Article 10 of the ECHR states that "this article shall not prevent States from requiring the licensing of broadcasting, television or cinema enterprises"[82]. This sentence of the Article 10(1) legitimises the State interference with media through licensing systems. The main objective of including this provision was the events that took place during the Second World War when different forms of media were misused for Nazi propaganda[83]. Accordingly, the overall conclusion could be that hate speech limitation to freedom of speech, as vague as the term is, still applies to the online world to the similar extent as in real life.

[81] Human Rights Committee General Comment 34, para. 43.
[82] Article 10 (1), ECHR.
[83] McGonagle, T. (2013). The Council of Europe against online hate speech: Conundrums and challenges. *Council of Europe Conference Expert Pape*r, 1–37., p. 7.

CHAPTER THREE

Theoretical framework

Theoretical framework underlying the interplay between freedom of expression and hate speech has been subject to contestations among scholars. Aside from the controversy regarding the scope of acceptable freedom of expression, it is crucial to understand that human rights and freedom of expression, in particular, were shaped and formulated in a particular time and context when the factor of social media did not exist. Compared to other expressionrelated issues, recent literature has rarely focused on a theoretical framework on how to regulate speech on Internet, including social media[84]. A considerable amount of literature has been published on the empirical data with a focus on the practical application of the international law and related problems, however, there is a deficit on the conceptualisation of the problem[85].

The classical theories regarding freedom of expression developed in the context where social media platforms and current communication mechanisms did not exist. Contemporary philosophy and democratic theory suggest revisiting the existing model. This chapter seeks to remedy these problems by providing a theoretical framework regarding online hate speech.

3.1. Theory of freedom of expression

Freedom of expression and its limitations have been subject too extensive legal, philosophical, and political debates and still remains contested. The theory of freedom of expression generally was widely discussed and the right was extensively interpreted in the second half of

[84] Wagner, B. (2016). Global Free Expression-Governing the Boundaries of Internet Content. *Springer International Publishing*, p. 12.
[85] *Ibid.*

the century[86]. The most prominent pieces of interpretations are quite contradictory towards each other.

Discussions regarding the doctrine of freedom of expression can be divided into three camps: those who provide arguments for knowledge and social progress (e.g. Mill), arguments for democracy (e.g. Meiklejohn) and arguments for personal autonomy (e.g. Rawls)[87].

John Stuart Mill, one of the most famous liberal defenders of free speech, considers any doctrine acceptable regardless of the immorality of the statement[95]. This approach suggests that public debate is a vital precondition for social progress[88]. Mill claimed that an individual can only enjoy their dignity fully if they are granted freedom of expression and consequently, he recognizes very few limitations to freedom of expression[89]. Therefore, the Millian concept does not take into consideration the consequences of freedom of expression, irrespective of their harm.

Meiklejohn, on the other hand, further interprets freedom of expression as a precondition for democracy[90]. From a political point of view, free speech provides a critical precondition for state institutions and is closely related to the political expression of ideas. He considered freedom of expression as a rather collective right that serves public purposes of democratic participation[91].

Rawls uses the argument of personal autonomy to justify the broad protection granted to freedom of expression[92]. According to him, freedom of expression is not only an instrument to achieve a higher public good, but bears importance on an individual level[93]. Therefore,

---

[86] *Ibid., p. 2.*
[87] Nash, V. (2013). Analyzing Freedom of Expression Online: Theoretical, Empirical, and Normative Contributions. *The Oxford Handbook Of Internet Studies* doi:10.1093/oxfordhb/9780199589074.013.0021 [95] Mill, J. S. (1978). On Liberty, edited with an introduction by Elizabeth Rapaport, p.16.
[88] *Ibid*.
[89] *Ibid*.
[90] Meiklejohn, A. (1965). Political Freedom: The Constitutional Powers of the People, New York: Oxford University Press.
[91] *Ibid*.
[92] Rawls, J. (1972). A Theory of Justice, New York: Oxford University Press.
[93] *Ibid.*

freedom of expression plays a crucial role in individual autonomy and applies not only political but any kind of expression.

Philosophers from the more recent period have focused on a different perspective on freedom of expression. For example, Dworkin, unlike most of the philosophers who primarily justify freedom of expression based on individual freedoms and liberty, argued that freedom of expression mainly serves the idea of equality[94]. Dworkin claims, that freedom of expression is a crucial right to equally be able to influence the environment that an individual is in. Therefore, every attempt to regulate freedom of expression would render the equality principle questionable[95]. For him, the idea of democracy is not a majoritarianism, but to facilitate equal participation.

However, Dworkin's claim has been the subject of criticism by those who find unregulated speech problematic in a number of cases, including hate speech. This kind of speech targets minorities and affects them in two ways. First, it silences them, and afterward, it subordinates them. Feminist scholar, MacKinnon argues that there is an inherent tension between equality and liberty[96]. According to MacKinnon, "the doctrine of free speech has developed without taking equality seriously - either the problem of social inequality or the mandate of substantive legal equality"[97]. Therefore, hate speech puts individuals in a position in which they are given unequal opportunities – in his own newspaper voices of the minorities are silenced and overshadowed by the majority's influence.

While minorities are theoretically offered an opportunity to defend themselves through counter-speech, the background of their speech is already damaged by the prior racist speech[98]. The oppressed have to defend themselves in an environment that is already biased by the

[94] Dworkin, R. (1996). Freedom's law : the moral reading of the American Constitution. New York ; Oxford : *Oxford University Press*, 1996.
[95] *Ibid.*
[96] MacKinnon, C. (1996) Only Words. Cambridge, MA: Harvard University Press, p. 71.
[97] *Ibid.*
[98] *Ibid*,. p. 31.

influence of the dominant culture. The oppressed and, therefore, unequal individuals are in an inferior position that undermines any possible counter-speech suggested by them.

MacKinnon further notes that "words and images are how people are placed in hierarchies" without giving them opportunities to live in an equal environment[99].

Philosopher Jeremy Waldron has similarly attempted to define the theory of freedom of expression less stringently and rely on individual dignity. He focuses on individual dignity and suggests that harms of hate speech directly attack the dignity of those who are the targets of such expression[100]. Accordingly, prohibition of such messages could contribute to the rights of excluded members of the society. Freedom of expression, at the same time, is portrayed as a tool not only for those who express views, but those, who are affected by the consequences.

3.2. Contemporary view

There are few scholars who specialize on the issue of freedom of expression in relation to modern communication changes. Professor Balkin is one of such scholars who, in 2004 suggested a different approach towards freedom of expression in relation to digital technologies, claiming that the social conditions of expression have significantly been changed[101]. Therefore, the theory of free speech should be shifted from a more traditional understanding of freedom of expression to a larger concept of promoting democratic culture[102]. Balkin argues that the free speech doctrine, adopted in the twentieth century, was a result of social conditions, such as the rise of mass media[103]. However, the conditions changed under the digital revolution at the beginning of the twenty-first century. Digital technologies

[99] *Ibid*., p. 31.
[100] Waldron, J. (2012). The harm in hate speech. Cambridge, Mass.; *London: Harvard University Press*, 2012, p. 96.
[101] Balkin, J. M. (2004). Digital speech and democratic culture: A theory of freedom of expression for the information society. *NyuL rev.*, 79, 1, p. 1.
[102] *Ibid.*
[103] *Ibid*.

do not only change the communication area but also create conflict over who controls informational capital[104]. The exercise of the right to freedom of expression strongly depends on the design of the technological infrastructure provided by Internet companies.

According to Balkin, the Internet era has raised the necessity to re-evaluate freedom of speech from a different perspective[105]. A similar reassessment was present when radio and television emerged. The digital revolution opened a way to the widespread cultural interaction that was not a case before. Despite the new opportunities created by the digital revolution, the dangers arise and therefore, it is necessary to accommodate properly[106]. Balkin describes how the digital age has changed the conditions of speech and demonstrates four key characteristics. The main characteristics of digital communication are that digital communications lower the costs of disseminating information and enable individuals to easily transcend geographical borders. Therefore, Balkin further underlines the dependence of free expression and infrastructure, resulting in merging those two[107].

The interplay between freedom of expression and internet-based platforms has further been explored by Wagner. According to him, the Internet, alongside other media and communications systems, constitutes a locus of power for states[108]. Not only does the Internet constitute a source of power, but it creates a common narrative in society[117]. Consequently, the control opportunities created by the Internet opens a door not only for control on an individual level but collectively[109]. Accordingly, there is a need for a revision of the approach regarding the freedom of expression in the era of social media.

[104] *Ibid*., p.2.
[105] *Ibid*. p.
[106] *Ibid*., p.3.
[107] *Ibid*.
[108] Wagner, B. (2016). Global Free Expression-Governing the Boundaries of Internet Content. *Springer International Publishing*, p. 5 [117] *Ibid*.
[109] *Ibid*., p. 16.

3.3. Summary of the theoretical framework

As it has already been illustrated, theoretical framework regarding freedom of expression has been shaped in a specific era with certain social conditions. However, the case of hate speech is unlike what it has been in the past. The key distinctive characteristic is that it occurs not only through individual interaction or mass media but via social media platforms.

The work of contemporary scholars suggests that the digital era challenges the traditional theory of freedom of expression due to a number of factors. It provides new means of interaction and reaching different groups of people as well as a possibility of control of information flaw. The theoretical approach needs to evolve in accordance with the specific nature of expression through social media. Furthermore, the debates regarding the hate speech online need to be revisited and adjusted to the new social condition. Otherwise, the gap between theory and practice could only further contribute to an inadequate understanding of the situation and underestimating the threats arising from it.

CHAPTER FOUR

Framing the problem - patterns of online hate

"It takes just one "friend of a friend" to infect a circle of hundreds of thousands of individuals with weird, hateful lies that may go unchallenged, twisting minds in unpredictable ways[110]*"*

Before proceeding to examine the effects of online hate speech, it will be necessary to look at the current forms and methods adopted by hate groups to disseminate ideas containing hate. A number of scholars point out that strategies of spreading online hate have evolved and surpassed the traditional expression of ideas by "ordinary people"[111]. The research of international organisations working on the issue demonstrates the significance and the scales of new forms of online hate[112]. In that respect, analysis of different historical factors and their influence on the development of the problem can be helpful.

4.1. Historical overview

The evolving nature of online hate speech is evident from the analysis of the development process of social media. A trend of rising hate speech needs to be considered in relation to a number of factors, such as the role of liberal perspective on freedom of expression in the development of social media in the twentieth century.

In the twentieth century, the leading source of information was the traditional media - frequently referred to as the "fourth estate", a public watchdog with the main role to shed light

[110] Foxman, A. H., & Wolf, C. (2013). Viral Hate: Containing Its Spread on the Internet. London: Palgrave Macmillan, p. 10.
[111] See below., p. 43.
[112] *Ibid*.

on government misconducts and abuses on power[113]. However, the information gathered by the traditional media did not reach to the same extent as the social media and was limited due to the commercial interests or other factors. Respectively the content was a subject to editorial supervision and relatively less complicated forms of judicial review[114].

The development process of traditional media forms was accompanied by the privatisation processes. At the beginning of the twenties century, in Western Europe, public space was primarily regulated by public institutions[115]. However, the consequent trend of economic liberalization affected every aspect of public/private distinction, including the public regulation of the media and lead to the privatisation of public institutions and services[116].

The western liberal tradition was also a foundation of the limits of information on the Internet where freedom of expression was seen as a fundamental human right that should not be interfered[117]. In the mid-1990s, global access to the Internet reached a significant level and became a part of everyday life and affected both - public and private spheres[118]. The Internet was created as a platform for communications and the *rationale* behind it was that this system would be difficult to control[119]. Lack of rules was considered to be a tool for democratization and less attention have been paid to the possibility of creating a sphere of chaos and lawlessness[120]. Due to egalitarian principles, the internet was designed to provide an opportunity for the free flow of knowledge, ideas, and information and accordingly, to be free

[113] Coe, P. (2015). Social Media Paradox: An Intersection with Freedom of Expression and the Criminal Law, *Information & Communications Technology Law*, (1), 16, p. 21.
[114] *Ibid*.
[115] Helberger, N., Kleinen-von Konigslow, K., & van der Noll, R. (2015). Regulating the New Information Intermediaries as Gatekeepers of Information Diversity. Info, 17(6), 50-71, p. 1.
[116] *Ibid*.
[117] Cohen-Almagor, R. (2015). Confronting the Internet's dark side: moral and social responsibility on the free highway. *Cambridge: Cambridge University Press* 2015, p. 5.
[118] *Ibid*.
[119] *Ibid*.
[120] *Ibid*.

from governmental intervention[121]. To ensure the achievement of these goals, various organizations have been established to promote freedom of expression on the Internet.

The overall picture regarding online hate speech changed significantly after the 2000s when social media rose. Social media platforms, nevertheless, have further changed the limits on information. Social media platforms are the Internet-based applications that enable every individual to share the content, such as ideas, photos, audio, and video files etc[122]. Social media is characterized by user-generated content, meaning that unlike traditional media that was described above, the content is not provided by a limited group of journalists or individuals, but everyone who has a purpose to publicly display their view. Therefore, communication through social media is restrained neither by the status of an individual who can share information nor by editorial or any other kind of prior supervision by appropriate bodies.

Characteristics of social media and its liberal foundations created an ideal venue for extremists to promote hate in the very early years of its creation. Online hate communications became more apparent and easy to spread. Individuals with far-right ideas have previously been isolated, limited with the physical space, however, the emergence of Web 2.0 opened a new venue for cooperation and as has been referred to by some scholars, created a "global racist subculture"[123]. After the introduction of Web 2.0, one of the first major "hate websites" - Stormfront emerged, which was created in 1995 by a former Ku Klux Klan leader[124]. The main objective of the webpage was to disseminate ideas of Neo-Nazism, White nationalism and White supremacy, advocating the initiation of "holy racial wars", resistance to immigration, and violent acts against minorities[125]. The US has further provided "safe haven"

[121] Banks, J. (2010). Regulating Hate Speech Online. *International Review of Law, Computers & Technology*, (3), 233.
[122] Cohen-Almagor, R. (2015). Confronting the Internet's dark side: moral and social responsibility on the free highway. *Cambridge: Cambridge University Press* 2015, p. 29.
[123] Banks, J. (2010). Regulating Hate Speech Online. *International Review Of Law, Computers & Technology*, (3), 233, p. 236.
[124] Gagliardore, I., D. Gal, T. Alves and G. Martinez. (2015). Countering online hate speech. *UNESCO Series on Internet Freedom*. Paris: UNESCO Publishing, p. 34.
[125] *Ibid.*

for hate groups that work to disseminate ideas such as Holocaust Denial and Christian identity. The number of users grow rapidly and reach new levels of activism, for instance, in 2016 the number of registered members of Stormfront reached around 300,000[126]. The Southern Poverty Law Center, one of the leading organisations focused on hate activity in the United States, has recorded the activities of 917 hate groups on social media in 2017[127].

4.2. Forms of online hate speech

Hate can be disseminated by both individuals and groups. Online hate speech authors do not have specific victims. Experience shows that the most common groups against whom such speech is directed are refugees and immigrants, as well as religious, ethnic minorities[128]. Such groups are marginalized and terrorist attacks and misconducts of individuals belonging to certain minority groups are frequently instrumentalised to undermine the public image of vulnerable people. This section will demonstrate different forms of online hate speech that exemplify the new methods of hate speech generally. The first one is a more traditional expression of intolerance towards different groups, while others are more recent, transformative measures that aim to change the perception of hate speech.

The major, commonly practiced form of hate is a direct expression of negative ideas towards target groups. In that case, users specifically address certain groups and disseminate ideas that show hostility towards them. This could be exemplified by the following cases: in 2010 violence was followed by the incitements of hatred by a public group on Facebook - "Kill a Jew Day"[129]. The group published the content encouraging violence against Jews, while the users commented that they could not "wait to rape the dead baby Jews"[130]. This is one example

[126] Cachia, A. (2014). A Web of Hate Rise of Online Abuse against Women, *The free Library*, *p. 56.*
[127] See: https://www.splcenter.org/issues/hate-and-extremism, (Last visited 22 May 2018).
[128] Klein, A. (2012). Slipping racism into the mainstream: A theory of information laundering. *Communication Theory*, 22(4), 427-448, p. 429.
[129] Citron, D. K., & Norton, H. (2011). Intermediaries and Hate Speech: Fostering Digital Citizenship for Our Information Age. *Boston University Law Review*, (4), p. 1435.
[130] *Ibid*.

out of tens of thousands of social media groups that are devoted to inciting hate[131]. Hate groups are not limited to the "traditional" victims of hate, such as Jews, black people, Muslims, homosexuals, but sometimes they target more unusual victims. For example, the Facebook group "Kick a Ginger Day" has encouraged physical attacks on students with red hair, and consequently, such attacks have happened[132]. Both of the pages were subsequently removed by the Facebook administration.

Nevertheless, modern hate groups have adopted a new, more innovative approach towards the incitement of hatred. A. Klein compares the process to money laundering, using the term "information laundering" to describe the newly fashioned form of hate speech[133]. Similarly to a money laundering system that allows criminals to disguise or conceal the results of their illegal actions, hate groups represent their hate-based information in as a form of knowledge[134]. According to Klein, the social media has created the ideal environment for hate groups to not only spread "toxic yet effective messages of cultural intolerance, racial superiority, or fear in a given society", but to edit, conceal or disguise facts and represent them as truth[135].

As a result, in the digital era, hate speech has reached new levels in more advanced and intellectual form. Such platforms provide ideal opportunities to hate groups not to simply spread hate messages, but to transform the understanding of hate and make their messages more justifiable. They replace hate speech with alleged scientific facts. Examples of such hate could be anti-Semitic groups disguised as Holocaust denial research organisations[136]. Through the new policy, hate groups aim to obtain more public confidence and legitimacy, change the

[131] *Ibid*.
[132] *Ibid., p.1437.*
[133] Klein, A. (2012). Slipping racism into the mainstream: A theory of information laundering. *Communication Theory*, 22(4), 427-448, p .431.
[134] *Ibid.*
[135] *Ibid., p.429.*
[136] *Ibid., p. 431.* [146]
Ibid.

perceptions of their victims as violent, hostile groups who have manipulated history and further rendered their grievances questionable[146].

As mentioned above, in recent years, social media, that is one of the trusted form of information, is not subject to any editorial filtering or criticism[137]. Therefore, through "information laundering", users can disguise anything - represent hate speech as a newspaper headlines to scholarly opinions[138]. As a result, the threshold of "trusted information" has increased. Despite the knowledge of the fact that social media lacks sufficient monitoring opportunities and that the information can be produced by any individual, public confidence in such posts and content is still high[139].

Lastly, the Hate Crime Prevention Institute has recognized another form of information laundering that is related to individuals pretending to be members of hate speech targets[140]. This scheme works in the following way: false pages pretend to be promoting, for instance, Muslim agendas, when in fact, they are publishing information that would cause a public outrage. Commonly, such pages or fake accounts support terrorism and violence. An example could be the events that took place (immediately) after the Lindt Cafe siege in Sydney in 2014, leaving four people dead[141]. In the aftermath of the attack, a number of pages on Facebook pretended to be local Muslims and expressed support for the attack. The pages were taken down after the Online Hate Prevention Institute reported them to both Facebook and the police, however, the damage was done - the post was seen by around 260 000 people, leaving a false impression regarding the Muslim community[142].

[137] *Ibid., p. 429.*
[138] *Ibid.*
[139] *Ibid.*
[140] Jakubowicz, A., Dunn, K., Mason, G., Paradies, Y., Bliuc, A. M., Bahfen, N., ... & Connelly, K. (2017). Cyber Racism and Community Resilience: Strategies for Combating Online Race Hate. *Springer*. p.58.
[141] *Ibid.*
[142] *Ibid.*

4.3. How hate speech works on social media

As noted above, social media platforms can be distinguished from other Internet web-pages as they are based on individual activism. Another name for this kind of connection is "web

2.0. applications", where these platforms are considered as "hosts", while they allow third parties to post content[143]. There are three key actors in online speech, including 1. the speaker, who sends information; 2. The listener, who receives information; 3. The intermediary or service provider who serves as a mediator between the speaker and the listener. Social media platforms act as an intermediary and provide communication between a speaker and a listener[144].

As demonstrated above, there are different types of hate speech. There are different standards of what kind of expression is lawful, some can be restricted and some must be restricted. Under those standards, it is a complex issue to decide what kind of hate speech is unlawful and therefore, a subject of restrictions.

As the Internet emerged, governments lost their power to control information gradually, however, not entirely. States developed certain techniques, what in the literature is referred to as new school speech regulation[145]. The new model of control of speech can be regarded as pluralistic, not dyadic[146]. To simplify the issue, unlike the dyadic model, there are minimum of three different groups of actors who participate in the process of speech - the "triangle includes" a/the state, on the one side, digital infrastructures that provide the means for communication on the second, and the speakers, on the third[147]. However, even though this system might seem triadic, the truth is that the new system of speech regulation is even more complicated. International organisations, such as Internet Corporation for Assigned Names

[143] Citron, D. K., & Norton, H. (2011). Intermediaries and Hate Speech: Fostering Digital Citizenship for Our Information Age. *Boston University Law Review*, (4), p. 1437.
[144] Sangsuvan, K. (2014). Balancing Freedom of Speech on the Internet under International Law. *North Carolina Journal Of International Law & Commercial Regulation*, 39701. P. 721.
[145] Balkin, J. M. (2018). Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation. *U.C. Davis Law Review*, 51(3) p.1187
[146] *Ibid*.
[147] Balkin, 2018, p. 1188.

and Numbers (ICANN) impose mandatory rules[148]. Moreover, expression of the end-users might be regulated by a number of private entities. For example, an unpopular speaker might be a user of social media, but a search engine company might "demote its page ranking"[149]. Furthermore, Internet Service Provider might also block the entire site. Companies, such as Facebook, YouTube, Twitter, and Google among them are the new governors of online expression.

It might be claimed that the same regime applied to mass media organisations in the twentieth century as their role was similar to the role of digital infrastructure providers. Radio and television broadcasters, newspapers, movie production studios, at first glance, provided a similar opportunity to individuals. However, their role cannot be equated to the role of social media companies. To start with, these companies produced their own content or published the content that was produced by a relatively small number of individuals[150]. By contrast, the new digital companies do not produce most of the content, rather, they simply serve as an intermediary for large audiences to share their views[151]. It is the public expression of the individuals that help big social media platforms to operate in the field.

Social media can be characterized by different qualities. First of all, the importance of market influence should be underlined. Even though social media companies offer their services to the users for free, they have adopted online advertising strategies that enable them to target users of a specific taste effectively and to modify their engines to advance their own interests[152]. Therefore, a financial condition of an individual or organization that is willing to advertise their page is decisive. It can also imply that the policies of the social media platforms depends on the popular view of the majority of the society to a great extent.

[148] *Ibid.*
[149] *Ibid.*
[150] *Ibid.*
[151] *Ibid*.
[152] Milosevic, T. (2016). Social Media Companies' Cyberbullying Policies. *International Journal Of Communication* (19328036), 105164-5185, p. 5167

Furthermore, information on social media can be collected and disseminated very quickly, irrespective of the geographical location of the speaker or the sender[153]. As mentioned above, social media platforms provide a unique opportunity for mass participation of anyone, who has a technical capability to connect to the Internet.

[153] Sangsuvan, K. (2014). Balancing Freedom of Speech on the Internet under International Law. North Carolina Journal of International Law & Commercial Regulation, 39701, p. 723.

CHAPTER FIVE
Critical analysis of the existing regulatory model

This chapter analysis whether and how hate speech on social media is addressed to protect members of the society. The following chapter is divided into two parts: first part analysis specific problems that are inherent to the existing model of hate speech regulations on social media first by looking at the different definitions of hate speech adopted by social media platforms and providing criticism of the practical application of those rules. The next crucial question is whether social media platforms are accountable with regards to human rights and the extent of their responsibility. Another important issue is the applicable law – since social media transcends borders, whose jurisdiction do they belong to. Finally, the first part analysis public/private dichotomy with regards to social media.

Second part of this chapter, on the other hand, provides more holistic analysis of the problem and aims to provide conceptualized understanding of the governance by social media.

5.1. Current mechanisms to address hate speech on social media

The question of who regulates content on social media arises the issue of identifying the actors who carry the main responsibility in protecting human rights. Accordingly, the central focus of this section will be to examine how hate speech on social media is globally regulated.

There is an agreement that social media platforms around the world are managed without central coordination and control[154]. There is no particular rule dealing with freedom of speech and especially, hate speech in relation to how to regulate communication between speakers and listeners. Social media regulation mechanism is often described as a "self-regulation" system. Self-regulation includes the creation and fulfillment of rules by the company itself,

[154] *Ibid.,* 734.

with very little state intervention or absence of such measures at all[155]. The contemporary regulation model seems to have disregarded the traditional "command and control" model through which states directly intervene against violation of freedom of expression[156]. Rather, the main actors are free in deciding what content constitutes illegal hate speech and, accordingly, what should be removed or blocked are social media platforms. In that framework, governments play a very limited role when it comes to communication technologies[157].

One of the rationales behind choosing self-regulatory mechanism over the traditional "command and control" model is its inability to keep up with fast-evolving technological progress[158]. The main purpose behind it is the impractical nature of government regulation both because of technological limitations and legitimacy concerns. Technical infrastructure to limit access to a particular content on social media is not available to states – they are only within the realm of individual social media companies[159]. Similarly, the issue of applicable legislation is problematic.

Another major obstacle related to social media regulation by private companies is the issue of legitimacy. Holding social media entities accountable raises a number of questions, such as what constitutes hate speech, what kind of hate speech need to be removed etc. Regarding those questions, it is arguable whether private actors can be competent enough to determine the answers. Thus, governance by private intermediaries carries specific implications on individual rights.

Although government's role in controlling the information on social media is very small, it does not mean that states are inactive and have completely delegated their power to private

[155] Milosevic, T. (2016). Social Media Companies' Cyberbullying Policies. *International Journal of Communication* (19328036), 105164-5185, p. 5168.
[156] Latzer, M., Just, N., & Saurwein, F. (2013). Self-and co-regulation: evidence, legitimacy and governance choice. *Routledge handbook of media law*, 373-397, p. 373.
[157] *Ibid*.
[158] Milosevic, T. (2016). Social Media Companies' Cyberbullying Policies. *International Journal of Communication* (19328036), 105164-5185, p. 5168.
[159] Shaw, L. (2011). Hate Speech in Cyberspace: Bitterness without Boundaries. Notre Dame Journal of Law, Ethics and Public Policy, 25(1), p. 281.

companies. Censorship is a frequently adopted measure by governments by either directly controlling internet infrastructure or by increasing pressures on intermediaries to meet certain requirements[160]. Alongside the private regulation, government interference might prove problematic.

The following subsections will focus on the drawbacks of the existing model and point out the major obstacles of regulating hate speech online. This part will first analyse the concept of hate speech that is suggested by social media, next – look at the legitimacy and accountability of social media and finally, identify problems related to the implementation of the relevant obligations.

5.1.1. Definition of hate speech by social media itself

Social media platforms are the primary entities that regulate freedom of expression of individuals, it does not indicate that they are in favour of complete anarchy and do not take measures to address the pressing issues. Policies are generally adopted regarding freedom of speech. For example, sexually explicit content, pornography, graphic videos showing someone being physically hurt, attacked or humiliated are a less controversial forms of expression that is regulated and limited by social media platforms[161].

Social media commonly addresses online hate speech through their own terms of service and community guidelines that in fact, are voluntary measures adopted by them[162]. Furthermore, these entities adopt their own individual definitions of hate speech either explicitly, or by providing a descriptive list of the terms related to it[163]. However, the voluntary efforts to determine the concept of hate speech can result in inconsistent and contradictory definitions.

---

[160] Balkin, J. M. (2017). Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation, 1149–1210. P. 1175.

[161] Sangsuvan, K. (2014). Balancing Freedom of Speech on the Internet under International Law. *North Carolina Journal of International Law & Commercial Regulation*, 39701, p. 740.

[162] Gagliardore, I., D. Gal, T. Alves and G. Martinez. (2015). Countering online hate speech. *UNESCO Series on Internet Freedom*. Paris: UNESCO Publishing, p. 29.

[163] See, for example, Facebook Community Standards, available at:

For example, YouTube's terms of service specifically address hate speech, as follows: "speech which attacks or demeans a group based on race or ethnic origin, religion, disability gender, age, veteran status and sexual orientation gender identity[164]". This definition encompasses a wide range of expression that does not necessarily constitute "incitement to discrimination, hostility or violence", as required under Article 20 of the ICCPR. The case of YouTube exemplifies a situation where the limitations provided by a private company are more restrictive than the requirements of the ICCPR.

Facebook, on the other hand has adopted community standards that provide a definition of hate speech that is similar to the scholarly definition of the term. Facebook community standards provide a prohibition of hate speech, specifically – verbal attacks and promotion of hatred based on the basis of race, ethnicity and national origin etc[165]. However, such prohibition does not have an absolute character and speech that can be considered hateful is allowed under certain circumstances, such as expression of humour/satire, social commentary, raising awareness etc[166]. The common requirement for allowing this kind of speech is that users use their authentic identity, through obliging page owners to indicate their identity.

The key problem with that arrangement is the possibility of creating tension between different international and national standards and the policies of social media. As private companies own the key global social media platforms, a question arises: whether these entities have human rights responsibilities to their users and more importantly, what are the consequences of conflict of different principles, such as, when those terms of service and community guidelines do not comply with international standards of freedom of expression.

The effects of voluntary definitions of hate speech by social media platforms can result in different forms. For example, a racist content published on Facebook by German users caused

https://www.facebook.com/communitystandards#hate-speech (Last visited 22 May 2018).

[164] YouTube Community Guidelines, available at: https://www.youtube.com/t/community_guidelines (Last visited 22 May 2018).

[165] Facebook Community Standards, available at: https://www.facebook.com/communitystandards#hate-speech (Last visited 22 May 2018).

[166] *Ibid*.

discontent in Germany in 2015. Users publish content about refugees containing hate speech and hoaxes representing refugees as criminals[167]. While some users reported the hate speech containing posts, claiming that it contradicted Facebook community standards, they were widely ignored by the administration[168]. Comments made by users urged violent acts against refugees, such as burning down refugee hostels[169]. Paradoxically, these kinds of comments did not violate Facebook's terms of service but constituted criminal acts in Germany. It required a personal engagement of public and private officials to solve the problem. In 2015 the German minister of justice met with Facebook managers to call for more efficient hate speech deletion process[170].

Thus, conflict between the definitions provided by social media platforms and national entities could not be solved in favour of governments without informal requests and personal engagement of public officials. The case illustrates that the decisive factor in such cases is private policies adopted by social media even if they do not fully comply with national legislation. Such regulation is problematic for its ability to circumvent a legitimate request of a government and thus, create regulation that is beyond a government control.

5.1.2. Identification of hate speech by private actors

Together with the definition of hate speech, another related challenge is the identification of exactly what constitutes such expression on social media, i.e. the practical application of the provisions. Online hate speech might be disseminated by individuals and organised groups that usually do not identify themselves as hate groups[171]. The common feature of these groups is that the description of hate pages contains a leading declaring the page against hate right

[167] Schieb, C., & Preuss, M. (2016). Governing hate speech by means of counter speech on Facebook. *In 66th ICA Annual Conference*, At Fukuoka, Japan (pp. 1-23), p.1.
[168] *Ibid*.
[169] *Ibid*., p. 2.
[170] *Ibid*.
[171] Cachia, A. (2014). A Web of Hate Rise of Online Abuse against Women, *The free Library*, p. 56.

before the statement that is a disguised hate speech[172]. Such groups have very strong identity views and promote a distinction between themselves and outsiders[183]. This kind of indeterminacy has been manipulated by some of the most famous hate groups, such as Ku Klux Klan (KKK) arguing that they represent the love of its own race and thus disguise their hateful expression[173].

The purpose of this strategy is to make it more difficult for the regulatory bodies of social media to quickly identify the hate speech and creates a precondition for bias[174]. This distinction serves a purpose to further hide the intent to stir up prejudices against others, where outsiders are often dehumanised and portrayed as a threat to their identity[186]. Furthermore, disguising intention of the content complicates the process of review as it is necessary to check the intention of the user and decide whether hate speech is used in a humorous/satirical way etc[175]. Moreover, such a difficult process can hardly be conducted by automatised means, as detecting hate speech and guaranteeing accuracy of removing such content would be questionable.

The amount of hate speech that is published on *Facebook* requires appropriate efforts and employment of labour. Another major obstacle is the language factor. A native speaker needs a certain amount of time to check if the complaint requesting the deletion of a post is wellfounded and goes against the community standards[176]. Additionally, understanding the context, social background is necessary to distinguish between real hate speech and humorous statements that sound ambiguous[189].

---

[172] Jakubowicz, A., Dunn, K., Mason, G., Paradies, Y., Bliuc, A. M., Bahfen, N., ... & Connelly, K. (2017). Cyber Racism and Community Resilience: Strategies for Combating Online Race Hate. Springer, p. 45 [183] Cachia, A. (2014). A Web of Hate Rise of Online Abuse against Women, *The free Library*, p. 56.
[173] *Ibid*. p. 56.
[174] Jakubowicz, A., Dunn, K., Mason, G., Paradies, Y., Bliuc, A. M., Bahfen, N., ... & Connelly, K. (2017). Cyber Racism and Community Resilience: Strategies for Combating Online Race Hate. Springer., p. 45 [186] Cachia, A. (2014). A Web of Hate Rise of Online Abuse against Women, *The free Library*, p. 56.
[175] Schieb, C., & Preuss, M. (2016). Governing hate speech by means of counter speech on Facebook. *In 66th ICA Annual Conference*, At Fukuoka, Japan (pp. 1-23), p. 2.
[176] *Ibid*., 2
[189]*Ibid*., 3

As a result, identification of hate speech by social media platforms is a difficult and time-consuming process. It requires adequate efforts and resources from social media platforms to thoroughly identify hidden messages and distinguish between speech that should be lawful and the one that should be prohibited. The specific nature of social media gives individuals almost endless possibilities to manipulate the content that only further aggravates the problem of identification of what constitutes hate speech.

5.1.3. Accountability issue

The role of social media in regulating hate speech, as demonstrated, is central. The key actors are private organisations "acting as a gateway for information and an intermediary for expression"[177]. Therefore, the questions on private authority and whether they should have responsibilities similar to public authorities are unavoidable. It is crucial to analyse to what extent social media platforms should be responsible for the protection individuals from exposing hate speech.

The key issue with the private regulation is that human rights obligations of non-state actors are very complex. There is a lack of academic consensus on the ways in which they can be held accountable. Theoretically speaking, states are the central duty-bearers, responsible for respecting, protecting and fulfilling human rights[178]. States are the ones who have the "monopoly of the legitimate use of physical force within a given territory"[179]. These obligations apply to the offline world as well. On the other hand, private companies, to the large extent, are not subject to international obligations[180]. Consequently, the question arises,

[177] Kaye, D. Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, (2016), A/HRC/32/38, para. 2.
[178] Weber, M. (1946). "Politics as a Vocation" In From Max Weber: Essays in Sociology, edited by Hans H. Gerth and C. Wright Mills, 77–128. Oxford, England: *Oxford University Press*, pp. 77–128.
[179] *Ibid.*
[180] Taylor, E. (2016). The Privatization of Human Rights: Illusions of Consent, Automation and Neutrality. *Global Commission on Internet Governance Paper Series,* p. 3.

why should any attention be paid to private actors at all. Some scholars even consider focusing on private entities as a distraction from the main problem[181].

The impact of multinational private actors has been recognized on the international level and resulted in the development of the UN Guiding Principles on Business and Human Rights (UNGPs)[182]. The abovementioned principles, although being non-binding, set down certain standards companies including social networking sites should comply with. In compliance with international human rights law, states need to reassure that business entities under their jurisdiction conform to the human rights law[183]. According to this document, all Internet intermediaries share a responsibility to respect human rights.

These principles, however, have rarely been implemented into practice by social media platforms[184]. They tend to adopt their own policies that question the balance between freedom of expression and other rights, such as dignity[185]. Notwithstanding, UNGPs do not have a binding nature and therefore, it does not give a conclusive answer to the question of private actors accountability towards the violations of human rights.

5.1.4. Applicable law/jurisdiction

As demonstrated above, states are not equipped with the technical capabilities to actually regulate freedom of expression on social media platforms, the only mechanism through which they control the flow of information is by blocking the entire web-page DNS[199]. Due to the

[181] *Ibid*.
[182] United Nations. (2011). Guiding principles on business and human rights: Implementing the United Nations "Protect, Respect and Remedy" framework. (A/HRC/17/31).
[183] *Ibid*., chap. I (A) (1).
[184] Gagliardore, I., D. Gal, T. Alves and G. Martinez. (2015). Countering online hate speech. *UNESCO Series on Internet Freedom*. Paris: UNESCO Publishing, p. 28.
[185] *Ibid*.
[199] MacKinnon, R., Maréchan, N., & Kumar, P. (2016). Corporate Accountability for a Free and Open Internet, (45)., p. 2.

global nature of social media platforms, the major obstacle in addressing hate speech is the issue of indeterminacy regarding the applicable law and jurisdiction.

International human rights law considers the concept of legality a vital precondition to limit freedom of expression. This is primarily evidenced in the requirements of Article 19 of the ICCPR. One of the key requirements of Article 19(3) of the ICCPR is the requirement of legal certainty, i.e. in order a restriction to be justified it needs to be "provided by the law". However, in the digital space, the requirement of legality becomes problematic as social media platforms do not operate within national boundaries. The grounds for limitation prescribed in law in one country might not constitute unlawful hate speech in another.

Major social media platforms, such as *Facebook*, *YouTube*, and *Twitter* etc. are US-based transnational corporations and therefore are mainly governed by US law[186]. Most of their data are processed and stored in the US[187]. The First Amendment of the US Constitution recognizes only a few limitations to freedom of speech and in that regard, provides wide protection[202].

The First Amendment does not protect hate speech only when the elements of imminent violence are present[188]. Moreover, the US has also attached reservations to Article 20 of the ICCPR that provides an additional layer of protection against hate speech[189]. The example of the US demonstrates the supremacy of national law - the constitution over international treaties.

Social media platforms generally take into consideration the requests of government authorities only in certain conditions, that is, when the case concerns criminal acts. They cooperate with national authorities when the case concerns criminal acts, as demonstrated by the transparency reports suggested by each of these actors[190]. Otherwise, social media

---

[186] Schieb, C., & Preuss, M. (2016). Governing hate speech by means of counter speech on Facebook. *In 66th ICA Annual Conference*, At Fukuoka, Japan (pp. 1-23), p. 2.
[187] Helberger, N., Kleinen-von Konigslow, K., & van der Noll, R. (2015). Regulating the New Information Intermediaries as Gatekeepers of Information Diversity. Info, 17(6), 50-71, p. 2. [202]*Ibid*.
[188] *Ibid*.
[189] *See, supra note,* 49.
[190] See, for example, Facebook Transparency Report, available at: https://govtrequests.facebook.com/ (Last visited 22 May 2018).

platforms tend to promote diversity and freedom of expression rather than obeying government requests.

Consequently, intention of the states to regulate hate propaganda on social media platforms is impeded by the issue of jurisdiction. Intention of a state to criminalise the publication of hate propaganda can be hindered by the absence of limits on geographical boundaries while publishing content on social media. Conflicts can and have occurred when states try to apply legislation extraterritorially into other jurisdictions. The case of *Yahoo!* in the early 2000s is an example of conflict of states who try to impose limits on the digital world[191]. The case of *Yahoo!, Inc v. La Ligue Contre Le Racisme et L'Antisemitisme* is a landmark case concerning two French organisations versus the Internet Service provider *Yahoo!*. The applicants claimed that Yahoo violated a French legislation that forbids the offering for the sale of Nazi memorabilia. Under French legislation, such behaviour is regarded as a serious crime[192].

The issue of jurisdiction and applicable law was problematic in that case. The content originated in the United States, and the ISP - *Yahoo*! was based in the US. However, the French court applied its own jurisdictional analysis and ruled that the company was liable under French legislation[193]. Taking into account the international character and the local impact of the content, the court imposed a financial sanction on *Yahoo!*.[194]

The subsequent events aggravated the question of jurisdiction even more. *Yahoo!* applied to the United States District Court, claiming that the enforcement of the decision of the French Court would violate the First Amendment of the Constitution[210]. The US Court agreed on the view of the applicant and found that even though hate speech might have serious negative consequences, it is still protected under the First Amendment unless it poses imminent threat

[191] La Ligue Contre La Racisme et L'Antisemitisme (LICRA) and Union Des Etudiants Juifs De France (UEJF) v. Yahoo! Inc. and Yahoo France (Paris, 2000)
[192] Okoniewski, E. A. (2002). Yahoo, Inc. v. LICRA: The French Challenge to Free Expression on the Internet. *American University International Law Review*, 18, 295.
[193] *Ibid.*
[194] *Ibid.*
[210] *Ibid.*

to its victims[195][196]. Therefore, enforcement of a foreign judgment that is inconsistent with the First Amendment would violate the protection of the Constitution.

The Yahoo case exemplifies the inherent cultural tensions regarding different conflicting views about freedom of speech. Even though nations are able to adopt regulations within their jurisdiction, they have difficulties applying rules extraterritorially. The decision that complies with the legislation of one state, can contradict another. Hate speech on social media, however, is an international issue, as its effects are not reduced to national borders. Consequently, the content that is produced in one country and has an effect on another continues to be unregulated.

5.1.5. Social media – public or private sphere

The classical understanding of freedom of expression is based on the public-private dichotomy[197]. The doctrine of free speech, accordingly considers the public sphere as a space where individuals are imposed liabilities, while the private sphere is a space for individual liberty[198]. More precisely, government regulations limit the freedom of an individual while without government interference, individuals are free to behave the way they consider appropriate.

Social media, however, puts the traditional public-private dichotomy under question. Online expression always happens through private intermediaries, including social media platforms. Increasing the role of online intermediaries is believed to affect civil liberties and democracy in various ways. One of the important theories is related to the privatisation-of-digital-publicsphere framework, as social media, that constitutes a public speech opportunities for the individuals, is being regulated by private rules, such as terms of service, community guidelines

[195] Yahoo!, Inc. v. La Ligue Contre le Racisme et L'Antisemitisme, 169 F. Supp. 2d 1181, 1186 (N.D. Cal.
[196])
[197] Feldman, S. M. (2016). Postmodern Free Expression: A Philosophical Rationale for the Digital Age. *Marq. L. Rev.*, 100, p. 1155.
[198] *Ibid*.

and standards, instead of laws, norms that are enforced by the legislative branches[199]. Therefore, it is not the government who determines the boundaries of freedom of expression in the public sphere, but social media companies.

Traditionally, it has been argued that there have always been privatized spaces in the offline world[200]. However, there are significant differences between online and offline private spheres. Private spaces in the offline world are strongly intertwined with specific geographical areas, while the online spaces transcend national borders and result in conflicting applicable jurisdictions and legislation.

5.2. Conceptualizing the problem

The following sections will be focused on providing more general conclusions regarding the drawbacks of private regulation with regards to hate speech on social media. This chapter further contrasts the effects of regulation by social media compared to the government regulation and explains the possible governance gaps that are created due to these arrangements.

5.2.1. What is wrong with private regulation by social media

Private governance means that social media entities govern the information. Social media plays a decisive role in constraining hate speech online. From a technical point of view, social media platforms are neutral actors as they are intermediaries and therefore, they do not create the content[201]. However, a major problem with this kind of perception is that in fact, social media platforms make day-to-day decisions regarding what content should be allowed on or

[199] Milosevic, T. (2016). Social Media Companies' Cyberbullying Policies. *International Journal of Communication,* 10(0), 22., p. 5167.
[200] *Ibid*.
[201] DeNardis, L., & Hackl, A. M. (2015). Internet governance by social media platforms. *Telecommunications Policy,* 39(9), 761-7706, p. 6.

removed from their platforms and under which conditions[202]. They determine the content of hate speech, how to identify it and how to react.

Debates regarding governance have caused a tension between those who advocate for greater government oversight and those in favour of self-regulation of the Internet. Governments, companies and non-state actors might pursue short-term agendas when it comes to internet regulation. As a result, scholars point out the existence of significant "governance gaps", that results in either a too permissive environment for private companies to contribute to violation of human rights or governments pressures on social media to impose limitations on freedom of speech .

The implications of private regulation conflict can be exemplified by a short movie called the "Innocence of Muslims" that was made in the US in 2012 and posted on YouTube. The movie was aimed to ridicule Islam as a religion and represent the Prophet Muhammad as an unreasonable, senseless man[203]. Within few days, the movie received worldwide attention and caused outrage in certain Muslim countries. Wide-scale protests emerged in Egypt, Libya, targeting the US for the role they played in creation and distribution of the video. The event is also associated with the Benghazi attack in Libya where the protesters killed the US Ambassador and three other Americans[204].

The case of "Innocence of Muslims" illustrates how social media has a bigger authority and control over the flow of speech than the governments. During these events, the government of the United States, the UN Secretary General - Ban Ki-Moon and Former Secretary of State

- Hillary Clinton strongly condemned the movie, claiming that it was "full of hate" and "disgusting and reprehensible" and further asked Google, which is YouTube's parent

[202] *Ibid*., p. 6.
[203] Sangsuvan, K. (2014). Balancing Freedom of Speech on the Internet under International Law. *North Carolina Journal Of International Law & Commercial Regulation*, 39701. p. 704.
[204] *Ibid*., p. 705.

company, to remove the content[205]. Google refused the request by claiming that the movie did not violate the terms of service and also, it did not constitute hate speech because it did not incite violence[221]. Google responded the existing situation by temporarily restricting access to the films in countries with a "sensitive situation", such as Libya, Egypt, etc.[222].

Consequently, freedom of expression and the right to equality – fundamental rights of an individual—are left in the hands of private actors. Private companies that undoubtedly play a significant role in terms of monitoring hate speech on social media, can disregard the requirements suggested by state officials and it can be claimed that such regulations leave gaps in governance with regards to the protection of human rights. Although it can be argued that social media sites usually provide certain safeguards through filtering techniques and user flagging, it would be difficult to conclude that such regulation provides comprehensive protection to individuals[223].

Furthermore, with regards to implementation, self-regulatory mechanism has other significant downsides, such as ineffective enforcement, less transparency, and accountability, and overall drawback - a possibility of prevailing private interests over the public ones[206]. One of the first difficulties related to the content filtering is the cost. Even though some social media platforms, such as *Facebook* and *Twitter* are large companies, the resources required for sufficient monitoring mechanisms are enormous and therefore, financial value of such monitoring is extremely high[207]. However, private companies frequently do not have sufficient motivation to interfere with the freedom of expression of their users and to limit hateful

[205] *Ibid*.
[221]*Ibid*., p. 739
[222]*Ibid*. [223]*Ibid.*

[206] Milosevic, T. (2016). Social Media Companies' Cyberbullying Policies. *International Journal of Communication* (19328036), 105164-5185, p. 5168.
[207] Wu, P. (2015). Impossible to Regulate? Social Media, Terrorists, and the Role for the U.N. *Chicago International Journal of Law,* 16(1), 281–311., p. 300.

speech[208]. Such high cost might be demotivating factor to social networking sites to act properly and address hate speech sufficiently.

Another crucial question is the source from where the policy of social media platforms derives from. The major online platforms are based in the United States and governed by their First Amendment free speech principles. However, the problem is that there is no guarantee that social media platforms will be based in the US or continue their current policy in the future. Examples of non-US based platforms are *VKontakte* (or *VK*) and *Odnoklassniki*, which are owned by Russian companies and have been blocked in certain states due to the possibility of Russian influence[209]. These platforms have a significant number of users compared to other major social media platforms in the world, for example, in Ukraine *VKontakte* had 11.9 million users while *Facebook* only around 8 million users[210].

Thus, in case these platforms gain global influence, there is no guarantee that they will not favour one view over the other. Moreover, roles might change in the future. In 2017, as a response to white supremacist demonstrations at Charlottesville, certain social media platforms blocked some neo-Nazi organisations[211]. There is no guarantee that the dominant social media players will continue to advocate human rights principles. Social media platforms can have a significant power generally, and if their commitments to human rights change in the future, there is little possibility to provide an adequate response at the international level.

[208] MacKinnon, R., Maréchan, N., & Kumar, P. (2016). Corporate Accountability for a Free and Open Internet, (45)., p. 2.

[209] Zhadova, M., Orlova, D. (2017). Computational Propaganda in Ukraine: Caught between external threats and internal challenges." Samuel Woolley and Philip N. Howard, Eds. Working Paper 2017.9. *Oxford, UK: Project on Computational Propaganda*, p. 8.

[210] *Ibid*.

[211] Balkin, J. M. (2018). Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation. *U.C. Davis Law Review*, 51(3), 1196

5.2.2. Government interference

Private regulation has specific features – although the infrastructure is owned by private entities, private regulation is not fully private. It is a common practice that states directly or indirectly pressure them to impose constraints[212]. Therefore, implications of the measures adopted by these private entities or the pressure imposed by state influence the protection of individuals from hate speech significantly.

Large multinational social networking sites increasingly find themselves constrained by governments, claiming their sovereignty over online flow of information[231]. A number of authors have investigated the attempts of repressive regimes to cut off communications during political unrests by blocking Domain Name Systems (DNS) and therefore, access to websites[213].

There are concerns regarding imposing obligations on social media actors to block and remove content. The restrictions apply prior to determination of the content legality by the Courts or any other form of supervision by government institutions[233]. Furthermore, there is no specific guidance provided for social media about the criteria of what constitutes hate speech[214]. The question arises whether these private entities are appropriate authorities to make a distinction between legal and illegal content and to provide difficult legal determinations.

The outcome of these measures could be excessive regulations adopted by social media to avoid financial sanctions or other kinds of responsibility imposed by states. Social media might not have another choice but to remove content that is, in fact, lawful, as a form of precaution[215]. As social media platforms have the role of gatekeepers in controlling content,

---

[212] *Ibid*., p. 1194. [231]
MacKinnon, R., Maréchan, N., & Kumar, P. (2016). Corporate Accountability for a Free and Open Internet, (45)., p. 2.
[213] DeNardis, L., & Hackl, A. M. (2015). Internet governance by social media platforms. *Telecommunications Policy*, 39(9), 761-7706, p, 2. [233]ARTICLE 19, Self-regulation and 'hate speech' on social media platforms, available at: https://www.article19.org/wp-content/uploads/2018/03/Self-regulation-and-%E2%80%98hatespeech%E2%80%99-on-social-media-platforms_March2018.pdf (Last visited 22 May 2018).
[214] *Ibid*.
[215] *Ibid*.

states have to ask them to intervene and block data[216]. One of the key challenges is related to the fact that regulation in not done primarily by the state. The issue has attracted the interest of various international figures.

In his 2016 report to the HRC, the Special Rapporteur on freedom of expression, David Kaye argued that states should not pressure private sector to unnecessarily interfere with freedom of expression through laws or other measures[217]. The report reiterated the idea that private intermediaries are not sufficiently equipped to determine the illegality of content. The possible dangers arising from such measures could be "over regulation"[218].

Collateral censorship is another term frequently used to describe a type of censorship where a state limits one's actions to control another one's speech[219]. Usually, those two actors do not represent the same enterprise. For example, holding a newspaper liable for its journalist's speech, it could be argued, does not constitute a significant problem[220]. Similarly, in the case of social media, collateral censorship affects the intermediary instead of those who produce speech.

Under certain circumstances, holding an individual for other person's speech is quite logical and does not contravene the idea of freedom of expression. To take one example, from the US Supreme Court case-law, Balkin demonstrates that holding a newspaper such as The New York Times for the speech of other people - its reporters or advertisers - is a well-established practice[221]. The Supreme Court of the US has found that responsibility applied to the newspapers equally in the judgment of *New York Times v. Sullivan*[222].

---

[216] DeNardis, L., & Hackl, A. M. (2015). Internet governance by social media platforms. *Telecommunications Policy*, 39(9), p. 6.
[217] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, 11 May 2016, A/HRC/32/38; para 40-44.
[218] *Ibid*.
[219] Balkin, J. M. (2018). Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation. *U.C. Davis Law Review*, 51(3), p. 1176.
[220] *Ibid*., p. 1176.
[221] Balkin, J. M. (2008). The future of free expression in a digital age. *Pepp. L. Rev.*, 36, 427, p.435.
[222] N.Y. Times Co. v. Sullivan, 376 U.S. 254 (1964).

At first glance, imposing sanctions on social media platforms for the content that has been published by end-users should be received in a similar manner. However, the claim about increased censorship is based on a different argument rather than a mere criticism of the idea of private actors being responsible for third parties expression. Newspaper or book publishers, media organisations have an increased interest in the work of their journalists – they have a vested interest to defend the expression of the speakers[223]. Unlike them, social media platforms are held responsible for the expression of end-users, who in the majority of cases, are less influential figures, and thus the interest of the intermediaries to guarantee their freedom of expression is low. Therefore, private intermediaries have a very limited incentive to defend speech of their users to prevent a lawsuit[224].

Collateral censorship takes place in the digital era through pressure of the states on private actors to censor, block and remove the speech of those individuals who use social media to express their views[225]. To do so, states adopt different strategies - they might impose criminal sanctions or penalties on the social media entities, they can also engage in "jawboning" - by urging them to do the right thing and tackle the content[226].

"Jawboning" is another form of government intervention – governments sometimes push platforms to censor the content on their platforms and such informal pressures and this method has worked in certain cases - social networking sites have blocked or deleted the content[227]. This kind of pressure appears as a voluntary action by the intermediaries, while in reality, it is not - states aim to achieve their goal by the circumvention of public law[248].

There are sufficient reasons to be sceptical about a government's attempts to intervene. "Jawboning" can become an illegitimate form of government action. Government has enormous resources to threaten private actors and use illegitimate measures, such as adopting

[223] Balkin (2008), p. 435.
[224] *Ibid*.
[225] Balkin, J. M. (2018). Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation. *U.C. Davis Law Review*, 51(3) p. 1177.
[226] *Ibid*.
[227] Bambauer, D. E. (2015). Against Jawboning. *Minnesota Law Review. 100*, 51, p. 1. [248]*Ibid*., p. 1.

a legislation that would affect the outcome[228]. Asymmetric power balance - private actors can only rely on their own resources, while the government has a greater capacity to influence on them[229].

Moreover, governments can bypass traditional formal checks and other processes such as the possibility of judicial review. The method of informal pressuring can be even more efficient than formal rule-making, since it poses uncertainty to the private actors leading to accept the general terms of government[251].

Consequently, what appears to be an issue of private regulation could change the entire understanding of freedom of expression. Professor Balkin makes a distinction between "old school" and "new school" speech regulations. While old school speech regulation is mainly directed at those who produce the content, new school speech regulations are not directed to the source itself, but the digital infrastructure that enabled them to become public[230]. The "old school" techniques regulating freedom of expression included imposing sanctions on those who produced the content, both - individual speakers and publishers, such as newspapers[231].

However, Balkin argues that under the "new school" speech regulation, responsibility has shifted towards social media platforms, which now are primarily responsible for the content that is produced by individuals[232].

Furthermore, Balkin provides the comparison to explain the difference between new and old schools of speech regulation and points out the shift of speech governance from dyadic to the pluralistic model[233]. In addition to private governance, the phenomenon of new school speech regulation gains particular importance. Digital governors and territorial governments' interests

[228] Bambauer, D. E. (2015). Against Jawboning. *Minnesota Law Review.*, *100*, 51., p. 5.
[229] *Ibid*., p. 6.
[251]*Ibid*., p. 6.
[230] Balkin, J. M. (2013). Old-School/New-School Speech Regulation. *Harvard Law Review*, (8), p. 2298.
[231] *Ibid*.
[232] *Ibid.*
[233] Balkin, J. M. (2018). Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation. *U.C. Davis Law Review*, 51(3), 1187.

are in conflict which results in a new type of regulation that is a new system of expression and governance. Regulation of online speech for states might be difficult as individuals might post anonymous messages, be located outside their jurisdiction and they might even not be humans at all, but bots[234]. Therefore, some states find it very convenient to move forward to new school speech regulation. Social media entities, on the other hand, are big corporations with an appropriate digital infrastructure to govern speech through blocking, filtering and removing content[235].

Balkin claims that in the twentieth century, when the international human rights treaties emerged and the free speech doctrine arose, the relationship between actors had a dyadic nature, as there were two actors - the state, on the one hand, and speakers and publishers, on the other[236]. The state regulated both - freedom of expression of speakers and publishers, by adopting legislation as a form of state censorship.

[234] *Ibid.*
[235] *Ibid.*
[236] *Ibid.*

CHAPTER SIX

Further proposals – the role of the international bodies

6.1. Regulation by the UN

Even though such debates have long existed and the harm caused by online mobilization of hate groups has been recognized, there has been little progress in collaboration at the global level[237]. Accordingly, such collaboration between different public and private actors will ultimately be necessary as hate speech has an impact not only on individual states, but on a global level and "unconstrained vitriolic hate speech contributes to declining civility both within and between nations"[238].

As social media transcends borders, scholars have discussed whether the proper body to regulate social media abuse by hate groups should be the UN. However, the recent attempts by the UN to communicate with the states to adopt a regulation regarding the Internet have not been successful, which suggests that this international body is unlikely to resolve the issue of social media regulation in the foreseeable future[239].

The issue of the social media governance is relatively new and unique, as it does not depend on state intervention. Certain models of global governance still can be traced in the modern international law. Yet, social media platforms need a source of justification for their model of governance. Wagner argues that similar to international organisations, such as the UN, Internet platforms require legitimacy to justify ongoing practices in the world[240]. Nevertheless, the

[237] Jakubowicz, A., Dunn, K., Mason, G., Paradies, Y., Bliuc, A. M., Bahfen, N., ... & Connelly, K. (2017). Cyber Racism and Community Resilience: Strategies for Combating Online Race Hate. Springer, p. 53.
[238] *Ibid*.
[239] See below.
[240] Wagner, B. (2016). Global Free Expression-Governing the Boundaries of Internet Content. *Springer International Publishing*, p. 158.

existence of institutions such as ICANN provides such a justification for the existing form of governance and creates an appearance of involvement of international bodies in the process[241].

Interestingly, Wagner develops the concept of "legitimacy theatre" to analyse the capacity of such international institutions[242]. The main argument for it is the fact that international institutions are very weak, without any real power to influence the outcome[243]. Accordingly, support for such mechanisms has its downside – not only does it provide insufficient guarantees of supervision, but also legitimises the existing system, where the *façade* of the regulation covers its real characteristics.

The inability of the UN to adopt a Convention on the internet regulation related-sphere was obvious at the World Conference of International Telecommunications (WCIT). Social media falls under the definition of "telecommunications" and is governed by the International Telecommunications Union (ITU) - an impartial organisation established in 1965 in charge of effective coordination of telecommunications networks[244].

Subsequently, in 2012, during the WCIT in Dubai, the internet governance issue was discussed[245]. The suggested outcome of the conference was to oblige every state to have an equal role in and responsibility for internet governance. However, the attempts were not successful as only 89 out of 144 states agreed to sign a non-binding act. The votes were split into two sides - countries such as Russia, China and other developing countries were in favour of the Final Act, while countries from Global West opposed it[246].

Aside from the controversy regarding geopolitical issues, the most contested part of the Final

---

[241] *Ibid*.
[242] *Ibid*. p. 159.
[243] *Ibid*.
[244] The International Telecommunication Union ('TILJ), U.N. NON-GOVERNMENTAL LIAISON SERVICE, http://www.un-ngls.org/spip.php?page=article-s&idartie=848 (Last visited 22 May 2018)
[245] Wu, P. (2015). Impossible to Regulate? Social Media, Terrorists, and the Role for the U.N. *Chicago International Journal of Law,* 16(1), 281–311., p. 294.
[246] See 2012 Final Acts of the WCIT, available at: https://www.itu.int/en/wcit-12/Documents/final-acts-wcit12.pdf (Last visited 22 May 2018).

Act was Resolution Plen/3, holding states equally responsible for international internet governance as the notion of equal responsibility can be related to certain costs[247]. While the US consistently argued that the Internet should not be regulated as it imposes limitations on free speech, Russia insisted an opportunity to have control over the Internet within its borders[270]. The polarisation of these positions reflect the long-standing difference in the understanding of freedom of expression, as the US has been in favour of unlimited speech while Russia consistently negated the importance of freedom of speech. Furthermore, from a political point of view, the greater government involvement in internet governance was seen as a threats to human rights[248].

In contrast, the Convention on Cybercrime, regulating cyber terrorism had been signed and ratified by the majority of the countries[249]. It may suggest that the UN can succeed in the limited convention, covering only social media. However, the significant differences between those two areas need to be taken into account. In terms of harm posed by hate speech on social media and cyber terrorism, they might have different incentives to counter it[250]. The negative outcomes of cyber terrorism are obvious, while hate speech is still only speech with less tangible outcomes. Therefore, cooperation among states regarding cyber-terrorism issues is much likely to take place in the future, compared to hate speech laws[274].

[247] *Ibid*.
[270] *Ibid*.
[248] *Ibid*.
[249] Wu, P. (2015). Impossible to Regulate? Social Media, Terrorists, and the Role for the U.N. *Chicago International Journal of Law,* 16(1), 281–311., p. 284.
[250] *Ibid*.
[274] *Ibid*.

6.2. Potential risks and benefits

The benefits of the UN involvement in social media regulation processes can be arguable. Two of the leading states who are permanent members of the Security Council[251], have significant problems regarding social media regulation. In the 2017 report of Freedom House,

China, and Russia have been referred to as pioneers in the use of "superstitious methods to distort online discussions and suppress dissent"[252]. Accordingly, the likelihood of cooperation among those states, let aside positive effects from such cooperation, is be highly doubtful.

The major risks that could hinder the process of international or regional cooperation among states is related to the unwillingness of coordination due to their national differences. Freedom of speech on social media almost inherently involves controversial attitudes regarding sensitive social interests and individuals. The reservations to the Article 20 of the ICCPR are a clear example of the controversy among states regarding the acceptable boundaries of freedom of expression in relation to hate speech[253].

Through enabling close connection between different states and different cultures, freedom of speech increases the possibility of some people being hurt by the expressions of others. Freedom of expression is regulated in a different manner in different jurisdictions. As noted above, the US legislation regarding hate speech is very broad, allowing the limitations over hate speech only when there is a threat of immediate violence while on the other hand, some European states criminalize defending, minimizing or denying the Holocaust[254]. Accordingly, challenges arise because of the absence of a consensus regarding hate propaganda, when it is completely legal in one state and constitutes a criminal offence in another.

---

[251] Security Council members, Available at: http://www.un.org/en/sc/members/ (Last visited 22 May 2018).
[252] Freedom House, Manipulating Social Media to Undermine Democracy, Available at: https://freedomhouse.org/report/freedom-net/freedom-net-2017 (Last visited 22 May 2018).
[253] Status of treaties, available at: https://treaties.un.org/Pages/ViewDetails.aspx?src=IND&mtdsg_no=IV4&chapter=4&clang=_en (Last visited 22 May 2018).
[254] Wu, P. (2015). Impossible to Regulate? Social Media, Terrorists, and the Role for the U.N. *Chicago International Journal of Law,* 16(1), 281–311., p. 294.

The problem would be more likely to exist even if states agreed to adopt the same approach regarding the regulation of hate speech. The content of speech is not always a decisive factor and could be dependent on the social conditions under which it is disseminated. As Shaw has noted, "the most destructive messages are those that rely on historically established hatred[255]". The major obstacle, aside from the willingness of the states, would be the fact that states have different social and historical context, therefore determining which words are harmful could be difficult. Speech directed against some minorities, such as Gypsies, could have a particular significance in a certain context of one state, while in another context it might not pose any threats at all[256]. In other words, the context under which hateful speech is interpreted is important and therefore, the impact of hate speech and its definition might vary from place to place.

Accordingly, the similar risks and benefits might derive in case of more regional cooperation. The European Union has already recognised the social harm caused by online hate speech. On 31 May 2016, the European Commission and social media companies (such as Facebook, Twitter, and YouTube) signed the document related to combating the spread of illegal "hate speech"[257]. As a result, the Code of Conduct on countering illegal "hate speech" online was adopted. The Code of Conduct focuses on "illegal hate speech", as defined in Framework Decision 2008/913/JHA of 28 November 2008 (Framework Decision), i.e. the "pubic incitement to violence or hatred directed against a group of persons or a member of such a group defined on the basis of race, colour, descent, religion or belief, or national or ethnic origin"[258] . Even though this can be considered as a voluntary effort of social media platforms to harmonize their terms of services, there still is a significant disagreement with respect to the precise meaning of "hate speech". Consequently, the role of regional organisations need to increase and become more harmonized.

---

[255] Shaw, L. (2011). Hate Speech in Cyberspace: Bitterness without Boundaries. *Notre Dame Journal Of Law, Ethics And Public Policy*, 25(1), 279-304, p. 287.
[256] *Ibid*.
[257] Portaru, A. (2017). Freedom of expression online: The code of conduct on countering illegal hate speech online. *Revista Romana de Drept European* 2017(4), 77-91, p. 78.
[258] Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law (OJ L 328, 6 December 2008, p. 55).

CHAPTER SEVEN

Concluding remarks

After the discussion provided above, it is crucial to look at the research questions and consider the answers. The research aimed to answer the question of what constitutes the major obstacles in regulating hate speech on social media; and consequently, what are the implications of the protection of individual rights by private intermediaries rather than by governments or global institutions. The answer to those questions, as illustrated above, is complex and depends on a number of factors.

A combination of two different problems poses challenges to the contemporary human rights regime. Firstly, hate speech has long been a problem affecting a wide circle of society, primarily minorities. Secondly, privatized regulation itself poses serious challenges not only with regards to hate speech, but issues such as privacy, pornography, etc. The combination of these two challenges creates an even more vague and ambiguous system that is difficult to address.

The first and primary problem is that the definition of hate speech is much contested. International and regional human rights treaties either provide vague terms or provide different levels of protection for expression. This can be illustrated by the divergence of definitions provided by the ICCPR, the CERD, and the CEDAW. Consequently, there is a risk of misapplication and misinterpretation of these rules not only by states but international organisations as well. The absence of a globally defined "hate speech" leaves the interpretation of this term in the hands of private actors or individual states.

Another challenge is addressing to online hate speech while respecting individual rights. There are serious concerns about individual freedom of expression on social media platforms. In that sense, private regulation leaves the protection of this right questionable, as its fulfillment depends on institutions that have neither the incentive(s) nor a strict obligation towards

society. Therefore, states have been criticised for an increased pressure on social media sites to limit the freedom of speech of individuals and provide a control mechanism.

A suggested solution to the problem is to focus on global institutions as private policies are not sufficient. It is necessary to draw attention to private intermediaries as the enjoyment of civil liberties has shifted and depends on them. Therefore, it is essential to understand the implications arising from private regulation and adopt better strategies to tackle the issue.

And finally, the facts provided in this thesis regarding the transformation of hate propaganda strategies demonstrate the urgency and necessity of an adequate response. Not only are more cooperative actions between states, global institutions, and private social media platforms necessary, but urgent and require an immediate response to avoid the enduring effects of hate speech on social media.

Bibliography

Table of cases:

A.W.R.A.P v. Denmark, the Committee on the Elimination of Racial Discrimination, no. 37/2006, 8 August 2007

La Ligue Contre La Racisme et L'Antisemitisme (LICRA) and Union Des Etudiants Juifs De France (UEJF) v. Yahoo! Inc. and Yahoo France, Paris, 2000

Malcolm Ross v. Canada, CCPR/C/70/D/736/1997, UN Human Rights Committee (HRC), 26 October 2000

N.Y. Times Co. v. Sullivan, 376 U.S. 254 (1964).Yahoo!, Inc. v. La Ligue Contre le Racisme et L'Antisemitisme, 169 F. Supp. 2d 1181, 1186, N.D. Cal. 2001

European Court of Human Rights (ECtHR)

Delfi AS v. Estonia (GC), Application no. 64569/09, (2015)
Erbakan v . Turkey, Application no. 59405/00 (2006)
Garaudy v. France, Application no. 65831/01 (2003),
Pavel Ivanov v Russia, Application no. 35222/04 (2007)
Seurot v. France, Application no. 57383/00 (2004)
Vejdeland and Others v Sweden, Application no. 1813/07 (2012)

Legal and semi-legal documents:

African Charter on Human and Peoples' Rights, (adopted 27 June 1981, entry into force 21 October 1986)

American Convention on Human Rights (adopted 22 November 1969, entry into force 18 July 1979)

Annotations on the text of the draft International Covenants on Human Rights, A/2929, 1 July 1955, paras. 189-194. Available at: http://www2.ohchr.org/english/issues/opinion/articles1920_iccpr/docs/A-2929.pdf (Last visited 22 May 2018).

Convention on the Elimination of All Forms of Discrimination against Women (adopted 18 December 1979, entry into force 3 September 1981).

Convention on the Prevention and Punishment of the Crime of Genocide (1948)

Council of Europe, Recommendation No. R 97 (20) of the Committee of Ministers to Member States on "hate speech", adopted on 30 October 1997, appendix.

Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law (OJ L 328, 6 December 2008, p. 55).s

General Comment No. 11: Prohibition of propaganda for war and inciting national, racial or religious hatred (Art. 20), 29 July 1983, the Human Rights Committee.

General Comment No. 15 of 23 March 1993, on article 4 of the Convention, the CERD Committee.

International Covenant on Civil and Political Rights (adopted 16 December 1966, entered into force 23 March 1976).

International Convention on the Elimination of All Forms of Racial Discrimination (adopted 21 December 1965, entry into force 4 January 1969).

Joint statement by the United Nations Special Rapporteur on freedom of opinion and expression, the OSCE Representative on freedom of the media and the OAS Special Rapporteur on freedom of expression on Racism and the Media, Available at: https://www.osce.org/fom/40053?download=true (Last visited 22 May 2018).
HR Committee, General Comment No. 34 on Article 19: Freedoms of opinion and expression, CCPR/C/GC/34, 12 September 2011.

Report of the Special Rapporteur on Minority Issues, Rita Izsák - Hate speech and incitement to hatred against minorities in the media, A/HRC/28/64, 5 January 2015.

Report of the Special Rapporteur to the Human Rights Council on Freedom of expression, states and the private sector in the digital age, A/HRC/32/38, 11 May 2016.

Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Promotion and protection of all human rights, civil, political, economic, social and cultural rights, including the right to development. A/HRC/36/1, 28 February 2008.

Study of the United Nations High Commissioner for Human Rights compiling existing legislations and jurisprudence concerning defamation of and contempt for religions, UN Doc. A/HRC/9/25, 5 September 2008.

The annual report of the UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression to the General Assembly, (2012).

The CERD Committee, Concluding Observations on Romania, CERD/C/ROU/CO/16-19, 13, September 2010.

The Rabat Plan of Action on the prohibition of advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence, A/HRC/22/17/Add.4, Appendix, adopted 5 October 2012.

Universal Declaration of Human Rights (adopted 10 December 1948).

Literature:

Balkin, J. M. (2004). Digital speech and democratic culture: A theory of freedom of expression for the information society. *NyuL rev.*

Balkin, J. M. (2008). The future of free expression in a digital age. *Pepp. L. Rev.*, 36.

Balkin, J. M. (2013). Old-School/New-School Speech Regulation. *Harvard Law Review*, (8), p. 2298.

Balkin, J. M. (2018). Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation. *U.C. Davis Law Review*, 51(3).

Bambauer, D. E. (2015). Against Jawboning. *Minnesota Law Review, 100*, 51.

Cachia, A. (2014). A Web of Hate Rise of Online Abuse against Women, *The free Library.*

Coe, P. (2015). Social Media Paradox: An Intersection with Freedom of Expression and the Criminal Law, *Information & Communications Technology Law*, (1), 16.

Citron, D. K., & Norton, H. (2011). Intermediaries and Hate Speech: Fostering Digital Citizenship for Our Information Age. *Boston University Law Review*, (4).

Cohen-Almagor, R. (2015). Confronting the Internet's dark side: moral and social responsibility on the free highway. *Cambridge: Cambridge University Press*.

DeNardis, L., & Hackl, A. M. (2015). Internet governance by social media platforms. *Telecommunications Policy*, 39(9).

Dworkin, R. (1996). Freedom's law: the moral reading of the American Constitution. *New York; Oxford: Oxford University Press*.

Feldman, S. M. (2016). Postmodern Free Expression: A Philosophical Rationale for the Digital Age. *Marq. L. Rev.*, *100*.

Foxman, A. H., & Wolf, C. (2013). Viral Hate: Containing Its Spread on the *Internet.London: Palgrave Macmillan*.

Gagliardore, I., D. Gal, T. Alves and G. Martinez. (2015). Countering online hate speech. *UNESCO Series on Internet Freedom*. Paris: UNESCO Publishing.

Hanzelka, J., & Schmidt, I. (2017). Dynamics of cyber hate in social media: A comparative analysis of anti-muslim movements in the Czech Republic and Germany. *International Journal of Cyber Criminology*, 11(1), 143–160. https://doi.org/10.5281/zenodo.495778

Helberger, N., Kleinen-von Konigslow, K., & van der Noll, R. (2015). Regulating the New Information Intermediaries as Gatekeepers of Information Diversity. Info, 17(6).

Jakubowicz, A., Dunn, K., Mason, G., Paradies, Y., Bliuc, A. M., Bahfen, N., ... & Connelly, K. (2017). Cyber Racism and Community Resilience: Strategies for Combating Online Race Hate. Springer.

Joseph, S. (2012). Social media, political change, and human rights. BC Int'l & Comp. L. Rev., 35.

Klein, A. (2012). Slipping racism into the mainstream: A theory of information laundering. *Communication Theory*, 22(4), 427-448.

Latzer, M., Just, N., & Saurwein, F. (2013). Self-and co-regulation: evidence, legitimacy and governance choice. *Routledge handbook of media law*.

MacKinnon, Catharine (1996) Only Words. *Cambridge, MA: Harvard University Press.*

MacKinnon, R., Maréchan, N., & Kumar, P. (2016). Corporate Accountability for a Free and Open Internet, (45).

McGonagle, T. (2013). The Council of Europe against online hate speech: Conundrums and challenges. *Council of Europe Conference Expert Pape*r.

Meiklejohn, A. (1965). Political Freedom: The Constitutional Powers of the People, *New York: Oxford University Press.*

Mill, J.S., (1978). On Liberty, edited with an introduction by Elizabeth Rapaport.
Milosevic, T. (2016). Social Media Companies' Cyberbullying Policies. *International Journal of Communication* (19328036), 105164-5185.

Nash, V. (2013). Analyzing Freedom of Expression Online: Theoretical, Empirical, and Normative Contributions. *The Oxford Handbook Of Internet Studies.*

Nowak, M. (2005). U.N. Covenant on Civil and Political Rights, CCPR Commentary, Kehl, *NP Engel Publisher.*

Okoniewski, E. A. (2002). Yahoo, Inc. v. LICRA: The French Challenge to Free Expression on the Internet. *American University International Law Review*, 18.

Portaru, A. (2017). Freedom of expression online: The code of conduct on countering illegal hate speech online. Revista Romana de Drept European 2017(4).

Rawls, J. (1972). A Theory of Justice, *New York: Oxford University Press.*

Sangsuvan, K. (2014). Balancing Freedom of Speech on the Internet Under International Law. *North Carolina Journal Of International Law & Commercial Regulation*.

Shaw, L. (2011). Hate Speech in Cyberspace: Bitterness without Boundaries. *Notre Dame Journal Of Law, Ethics And Public Policy*, 25(1), 279-304.

Siegel, M. L. (1998). Hate Speech, Civil Rights, and the Internet: The Jurisdictional and Human Rights Nightmare [comments]. Albany Law Journal of Science & Technology, (2).

Taylor, E. (2016). The Privatization of Human Rights: Illusions of Consent, Automation and Neutrality. Global Commission on Internet Governance Paper Series.

Wagner, B. (2016). Global Free Expression - Governing the Boundaries of Internet Content. *Springer International Publishing*: Imprint Springer.

Waldron, J. (2012). The harm in hate speech. Cambridge, Mass.; *London: Harvard University Press*, 2012.

Williams, M. L., & Burnap, P. (2016). Cyberhate on Social Media in the Aftermath of Woolwich: a Case Study in Computational Criminology and Big Data. *British Journal of Criminology*, 56(2).

Zhadova, M., Orlova, D. (2017). Computational Propaganda in Ukraine: Caught between external threats and internal challenges." Samuel Woolley and Philip N. Howard, Eds. Working Paper 2017.9. *Oxford, UK: Project on Computational Propaganda*.

Online resources:

2012 Final Acts of the WCIT, available at: https://www.itu.int/en/wcit-12/Documents/finalacts-wcit-12.pdf (Last visited 22 May 2018).

ARTICLE 19, (2015) 'Hate Speech' Explained: A Toolkit. Available at: https://www.article19.org/data/files/medialibrary/38231/'Hate-Speech'-Explained---A-Toolkit%282015-Edition%29.pdf (Last visited 22 May 2018).

ARTICLE 19, Self-regulation and ‘hate speech’ on social media platforms, available at: https://www.article19.org/wp-content/uploads/2018/03/Self-regulation-and-%E2%80%98hate-speech%E2%80%99-on-social-media-platforms_March2018.pdf (Last visited 22 May 2018).

CoE Factsheet 'Hate speech' (2012), available at: https://www.echr.coe.int/Documents/FS_Hate_speech_ENG.pdf (Last visited 22 May 2018)

Facebook Transparency Report, available at: https://govtrequests.facebook.com/ (Last visited 22 May 2018).

Fear and loathing, available at: https://www.theguardian.com/uk/2004/aug/12/race.world (Last visited 22 May 2018).

Freedom House, Manipulating Social Media to Undermine Democracy, Available at: https://freedomhouse.org/report/freedom-net/freedom-net-2017 (Last visited 22 May 2018).

Online Hate Prevention Institute. 2014. “Press release: Launch of an online tool to combat hate”. Available at: http://ohpi.org.au/press-release-launch-of-online-tool-to-combat-hate/ (Last visited 22 May 2018).

Security Council members, Available at: http://www.un.org/en/sc/members/ (Last visited 22 May 2018).

Status of treaties, available at: https://treaties.un.org/Pages/ViewDetails.aspx?src=IND&mtdsg_no=IV4&chapter=4&clang=_en (Last visited 22 May 2018).

The International Telecommunication Union ('TILJ), U.N. NON-GOVERNMENTAL LIAISON SERVICE, http://www.un-ngls.org/spip.php?page=article-s&idartie=848 (Last visited 22 May 2018).

The Southern Poverty Law Centre Hate Map, Available at: https://www.splcenter.org/hatemap (last visited 22 May 2018).

YouTube Community Guidelines, available at: https://www.youtube.com/t/community_guidelines (Last visited 22 May 2018).